Circuit and CAD Techniques for Expanding the SRAM Design Space

PhD Proposal

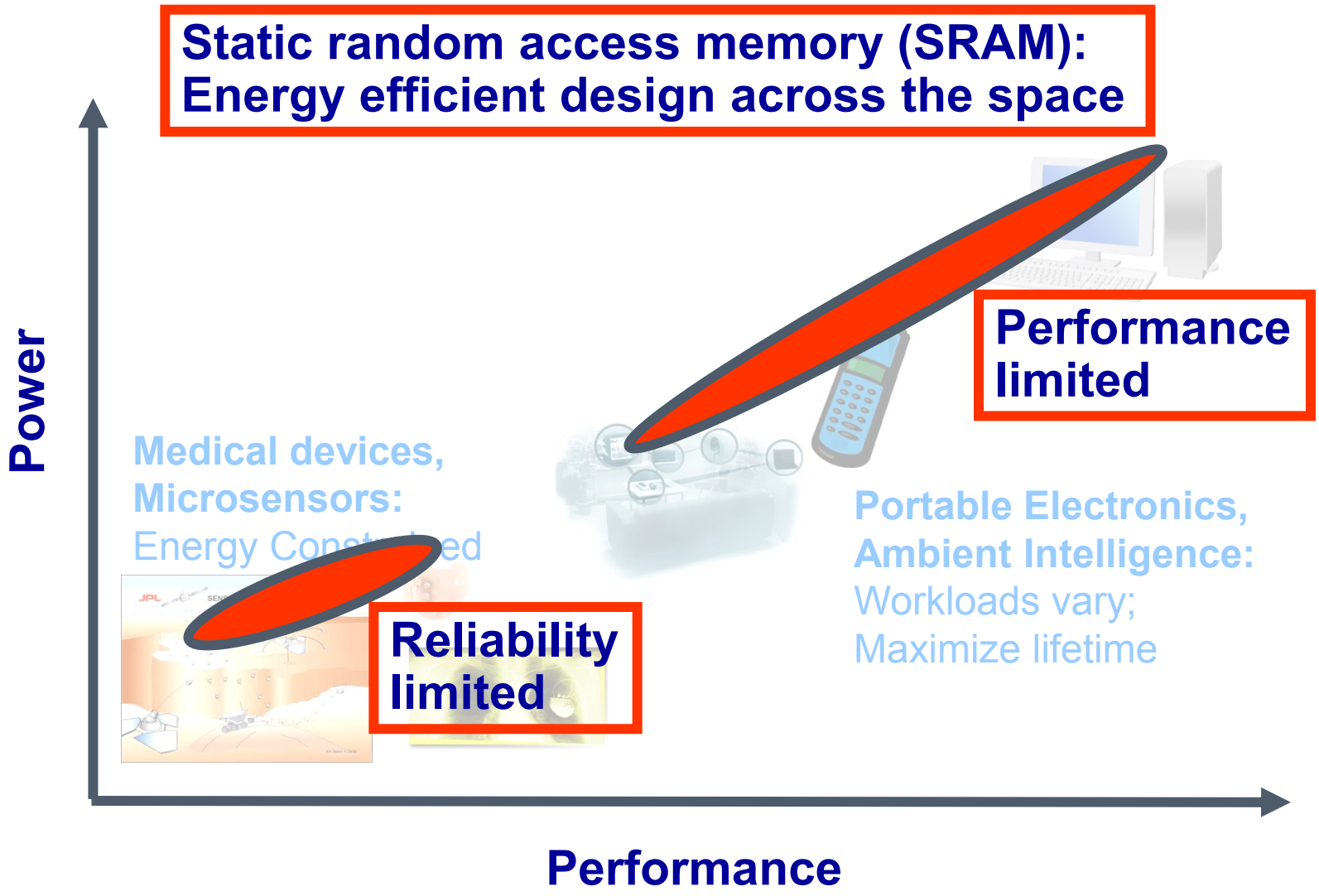
Jim Boley

April 24, 2013

Robust Low Power VLSI

Power Limits Across the Application Space

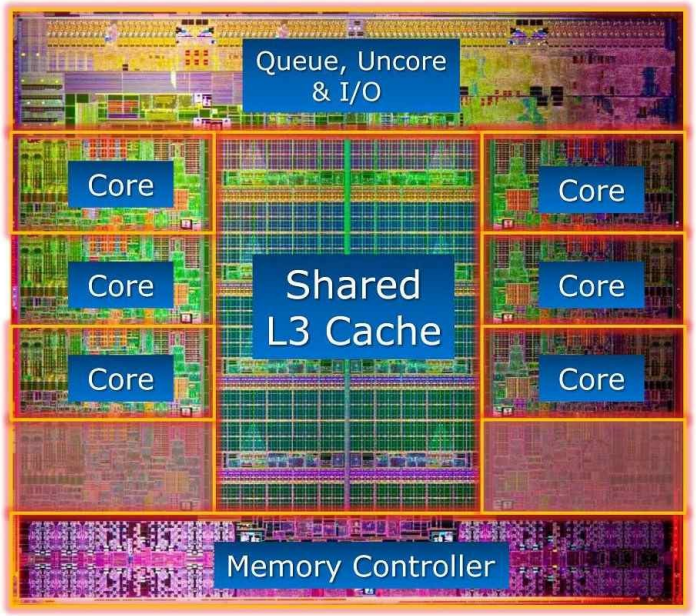
Energy Efficiency Across the Application Space



Why Focus on SRAM?

- High speed, high density memory option
- Consumes large amounts of die area
- High leakage energy
- Often sets minimum operating voltage (V_{MIN})
- Memory access → performance bottleneck

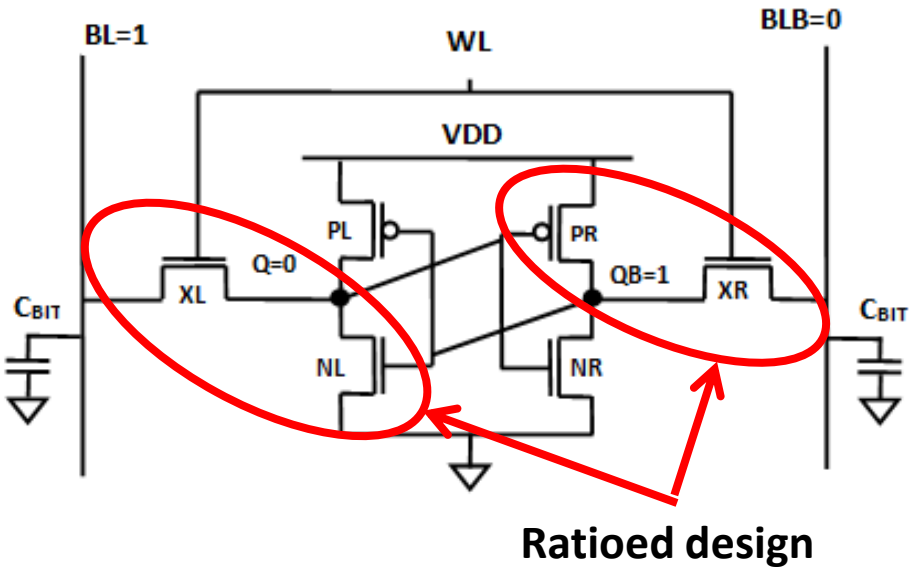
Intel® Core™ i7-3960X Processor Die Detail



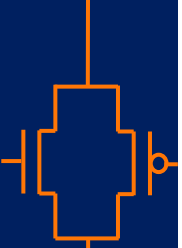
www.intel.com

Key Challenges

1. Read static noise margin (SNM)
2. Write-ability
3. Read access failures
4. Evaluating Yield
5. Evaluating design decisions



Ratioed circuit design coupled with close to minimum sized devices leads to a higher sensitivity to variation



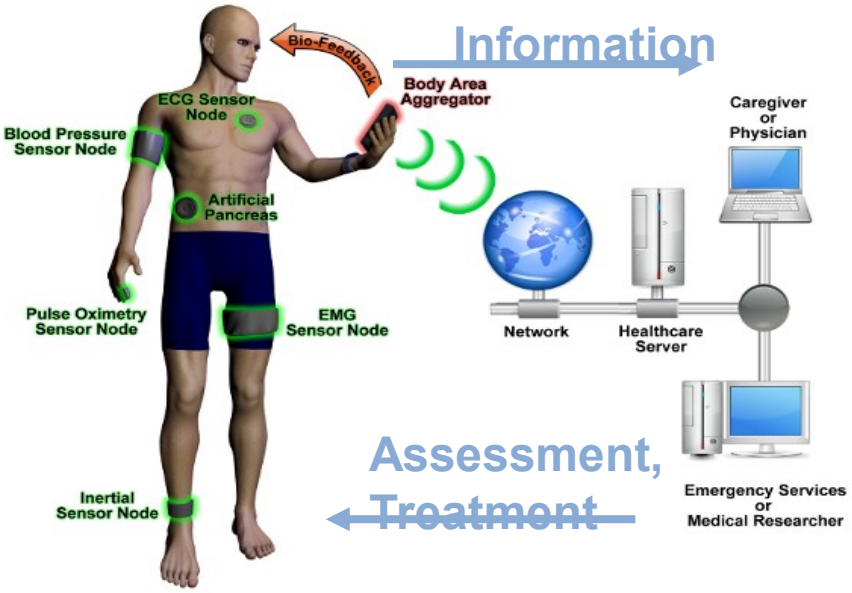
Thesis Statement

By developing a set of methods and tools for evaluating SRAM design tradeoffs, we can push SRAM designs to lower operating voltages, increase yields, and evaluate both circuit and architectural design decisions.

Outline

- Motivation
- **Sub-threshold SRAM design**
- Analysis of sub-threshold bitcell and assist methods
- A method for evaluating dynamic V_{MIN}
- Extending the capabilities of ViPro
- A canary based feedback system for PVT tracking
- Schedule

Motivation for Sub-V_T in Body Sensor Networks



- Input signals are low frequency (low clock speed)
- Highly energy constrained
- Goal: run solely off harvested energy (10's of µWs)

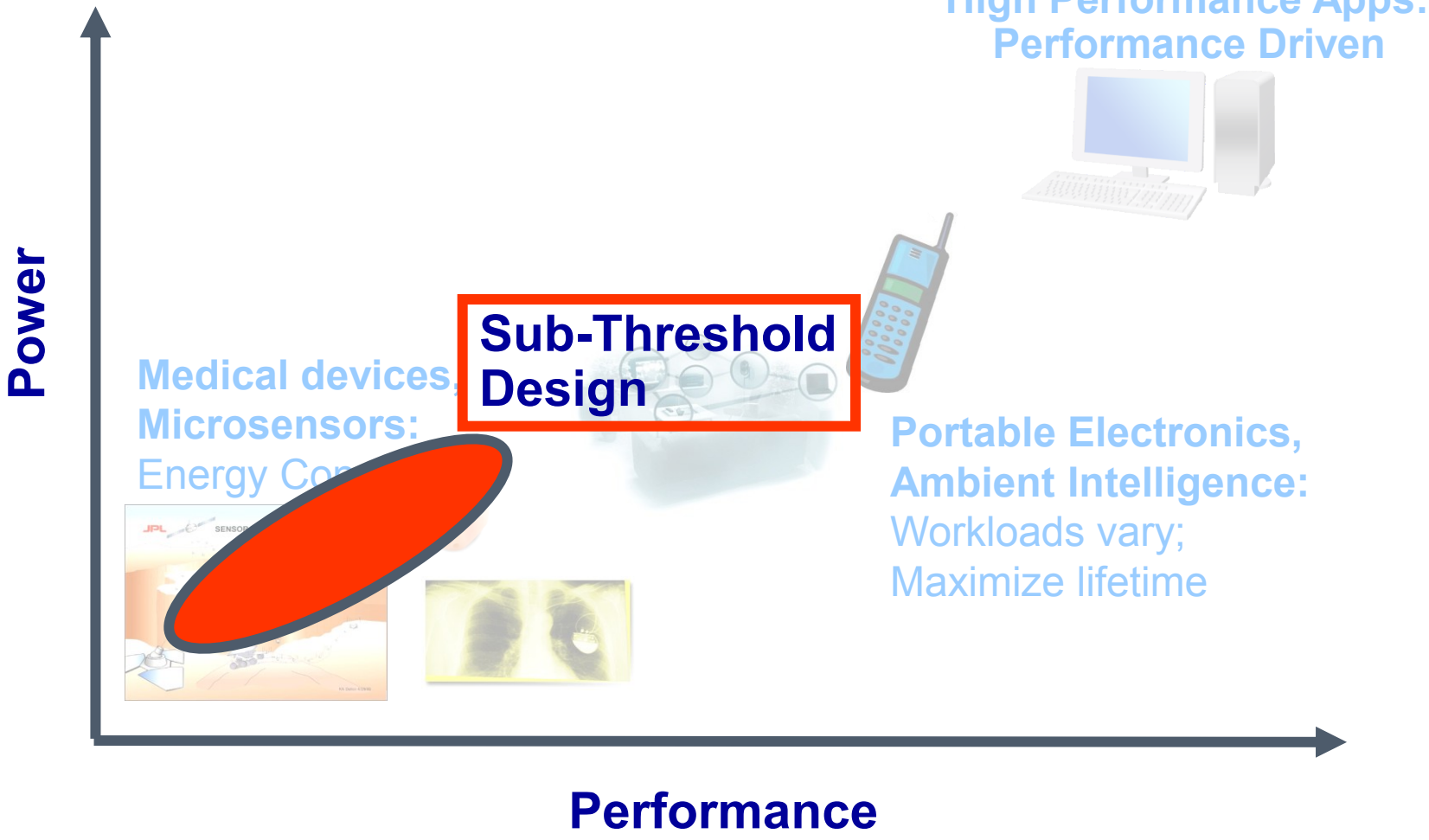
Sub-threshold operation minimizes energy per operation

Challenges: Sub-Threshold SRAM

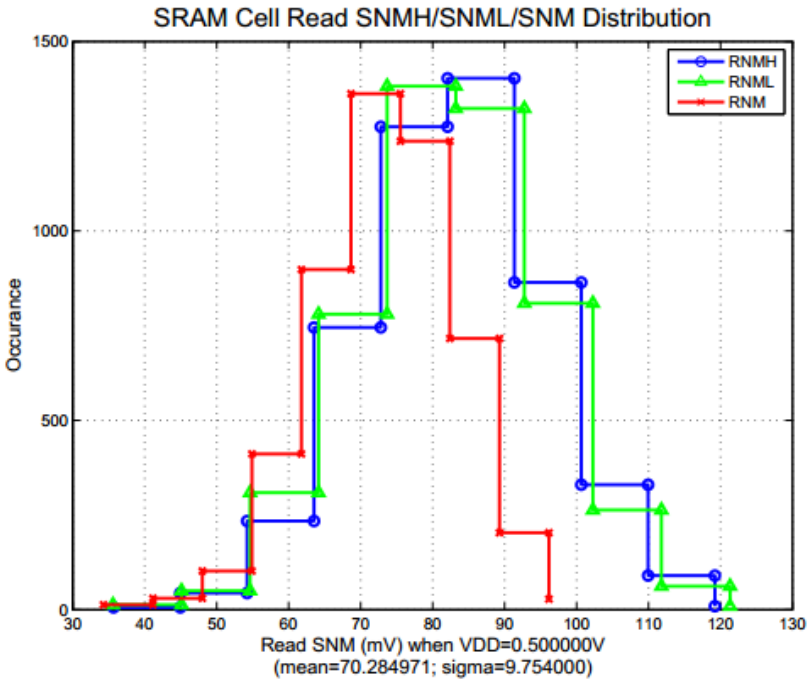
- Familiar Problems
  - Hold SNM, Read SNM, Write NM
- New Problems:
  - Reduced Ion/Ioff ratio (Read access failure)
  - Exaggerated V_T variation impact (Ion varies exponentially with Vt)
- Possible Solutions
  - Read SNM: Use a read buffer to decouple the data
  - Write NM: Use write assist (Boosted WL/Negative BL VSS)

Research Question:
How can we design an ultra low power SRAM capable of reliable operation at an operating voltage of 500 mV and a frequency of 200 KHz?

Scope

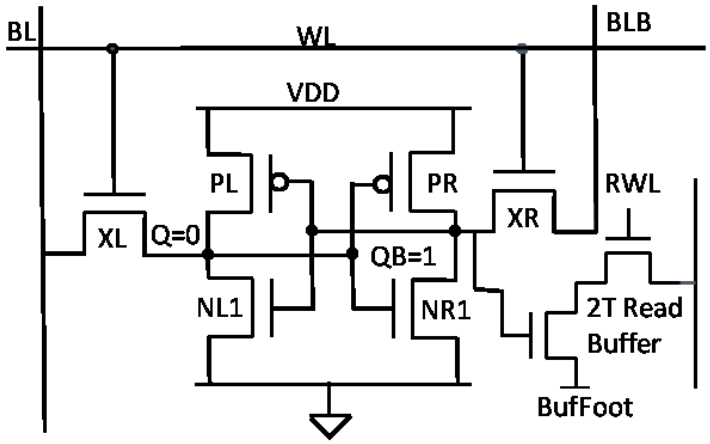

High Performance Apps:
Performance Driven
Sub-Threshold
Design
Medical devices,
Microsensors:
Energy Co
Portable Electronics,
Ambient Intelligence:
Workloads vary;
Maximize lifetime
Power
Performance

Approach: Addressing Read SNM

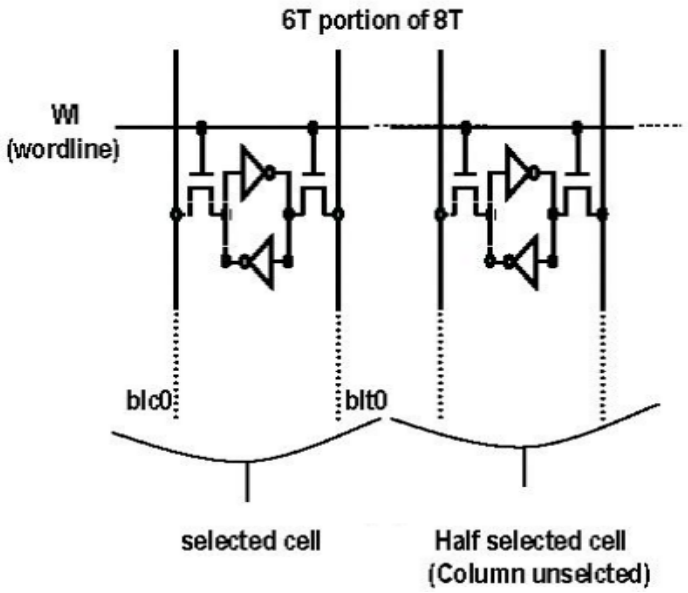


- Worst case SNM 35 mV
- 8T decouples data from the read operation
- **New problem: half select stability**

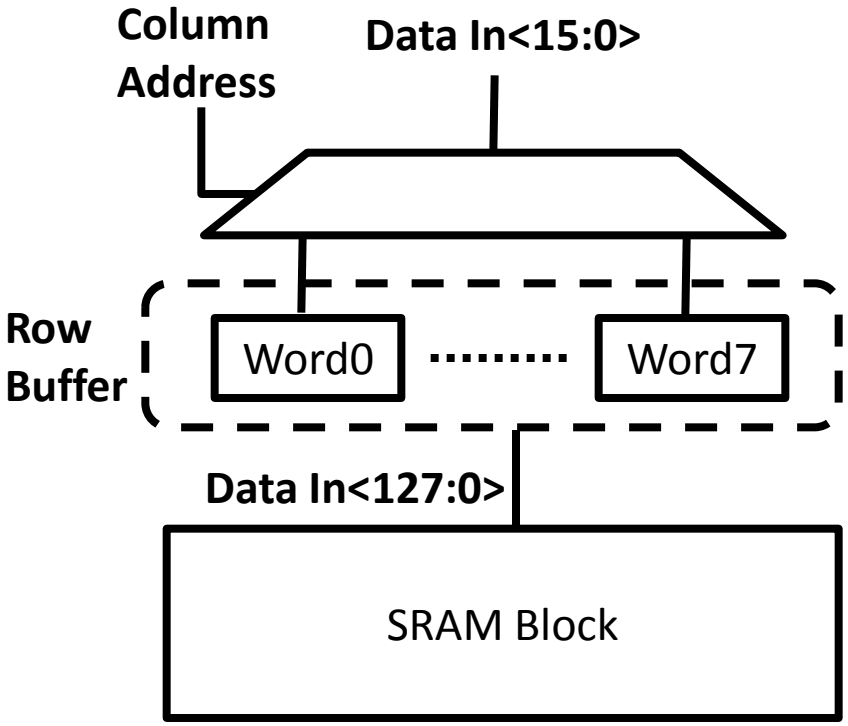
Use a buffer to fix Read SNM



Approach: Addressing Half Select



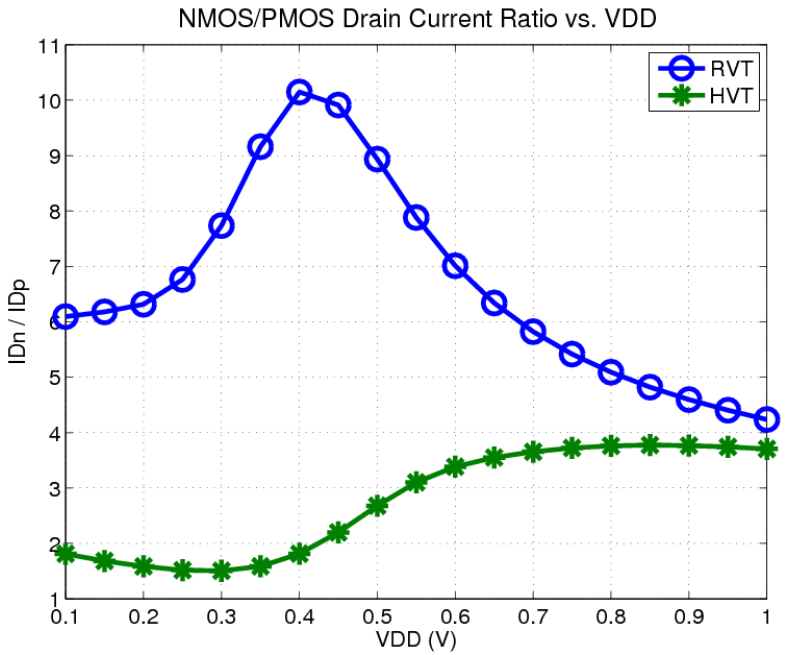
[R. Houle, VLSI Circuits 2007]

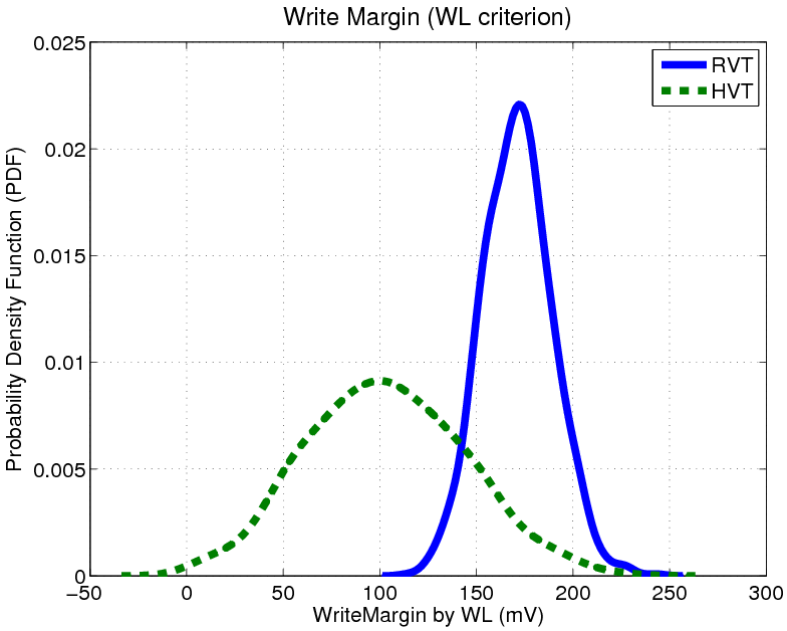


- Both data and instruction memory used as FIFO
- Memory written once every 8 cycles

Use a row buffer to prevent half select instability

Approach: Addressing Write-ability

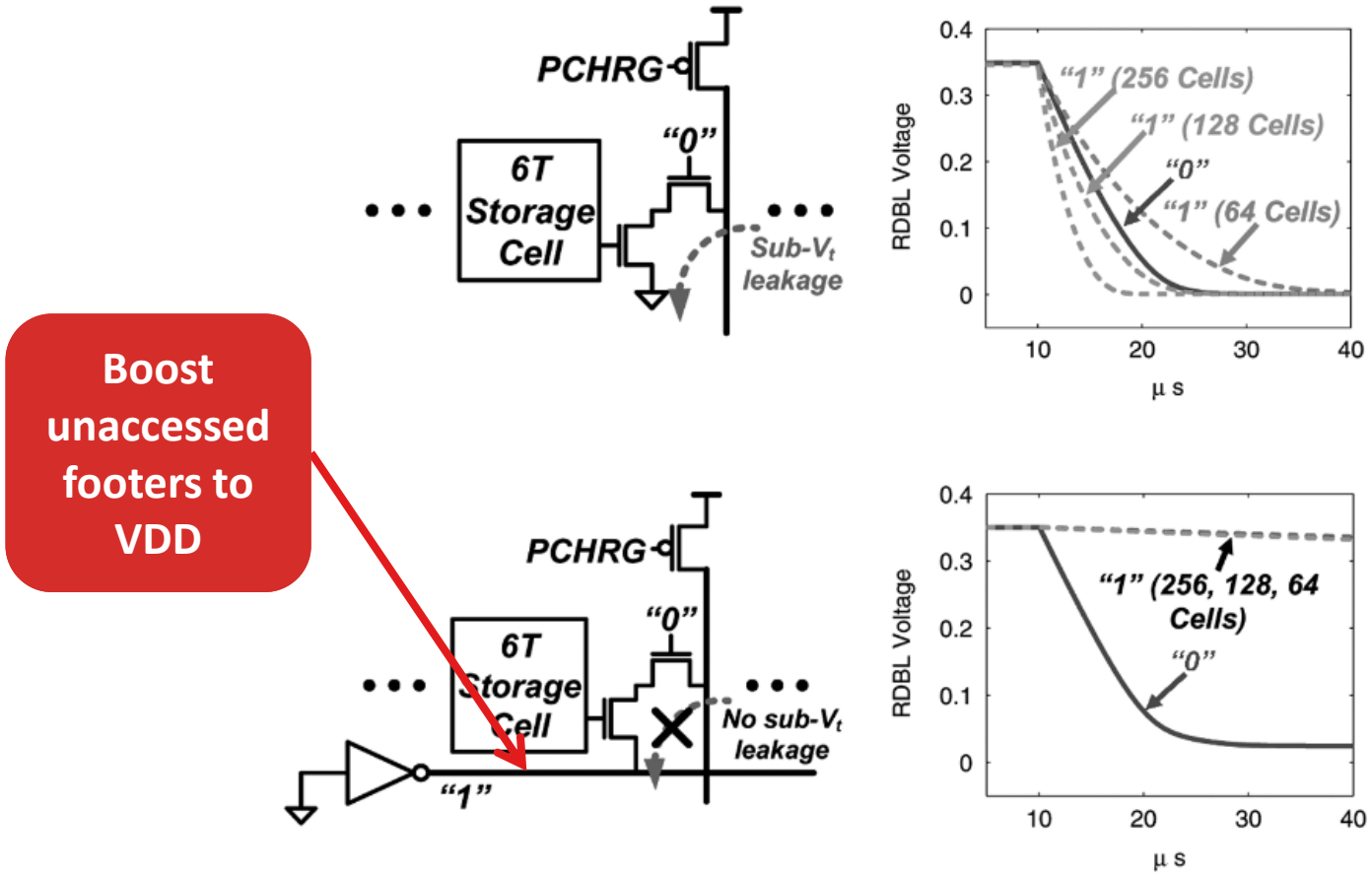




- Write-ability dependent on N/P current ratio
- RVT devices → high N/P ratio in sub-vt
- Downside: 22x increase in leakage per bitcell

Use regular VT devices to ensure write-ability

Approach: Addressing Read Access



[N. Verma, JSSC 2008]

Boost read footer to prevent read access failures

Metrics for Evaluation

1. Minimum operating voltage at which reliable operation is achievable
   - Success: reliable operation at 500 mV and 200 KHz
2. Total energy per access
   - Success: 10's picojoules range

Contributions

- Methodology for designing embedded Sub-V_T SRAMS
- Reliable operation from 0.3V – 0.7V at 200 KHz
- Energy per access → 12.1 pJ
- Leakage per cycle → 6.6 pJ

	This Work	[5]	[22]	[6]	[23]	[7]
Sensors	ECG, EMG, EEG	ECG	Neural, ECG, EMG, EEG	EEG	ECG, TIV	Temp, Pressure
Supply Voltage	30mV, -10dBm	1.2V	1V	1V	1.2V	0.4V/0.5V
E Harvesting	Thermal, RF	×	×	×	×	Solar
Memory	5.5kB (0.3V-0.7V)	42kB (1.2V)	×	×	20kB (1.2V)	5kB (0.4V)

[Zhang, JSSC 2013]

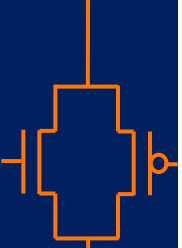
Future work: reduce leakage per cycle by using high V_T devices

Outline

- Motivation
- Sub-threshold SRAM design
- **Analysis of sub-threshold bitcell and assist methods**
- A method for evaluating dynamic V_{MIN}
- Extending the capabilities of ViPro
- A canary based feedback system for PVT tracking
- Schedule

Motivation for Sub-threshold SRAM

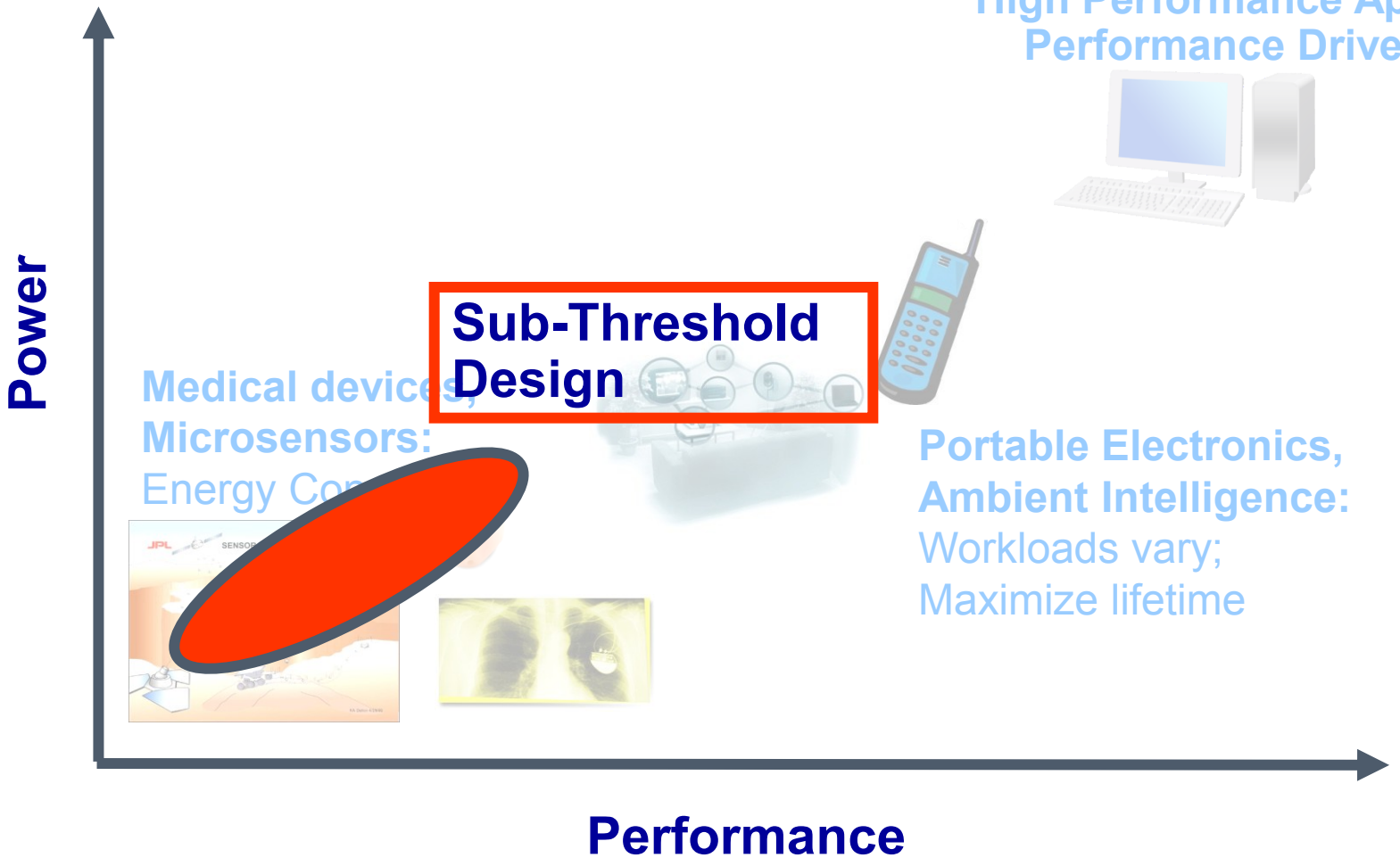
- Motivation:
  - Reduction in leakage power
  - Quadratic active energy savings
- Challenges:
  - Increased sensitivity to variation
  - Reduced Ion/Ioff ratio
  - Read Static Noise Margin
  - Write-ability
  - Read access stability
- **Approaches:**
  - **Alternative bitcell designs**
  - **Read and write assist methods**



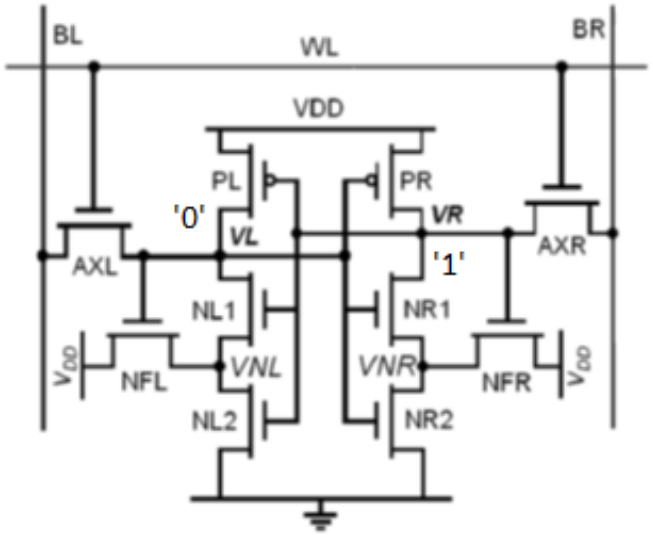
Hypothesis

By testing different combinations of bitcell topologies and assist methods, we can determine which approach results in the largest reduction of read and write V_{MIN} over the nominal case (6T bitcell with no assist methods).

Scope

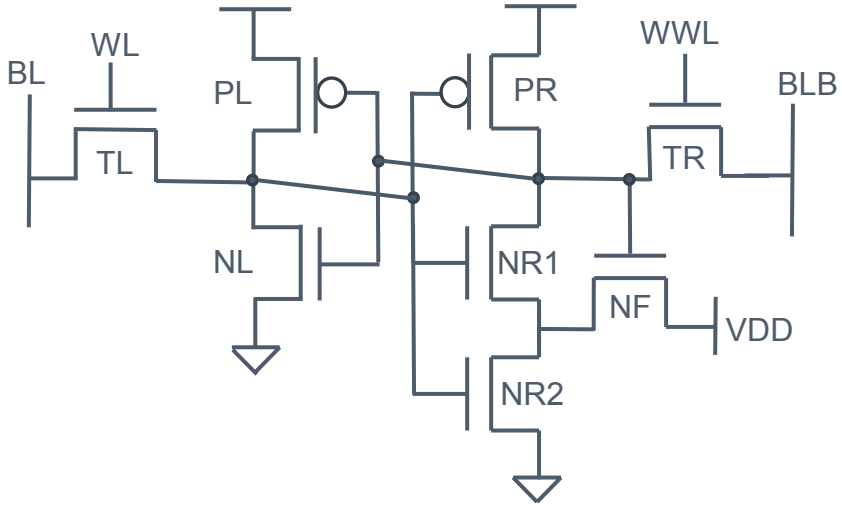

High Performance Apps:
Performance Driven
Sub-Threshold
Design
Medical devices;
Microsensors:
Energy Con
Portable Electronics,
Ambient Intelligence:
Workloads vary;
Maximize lifetime
Power
Performance

Approach:Non-6T Cell for Read Stability



Schmitt-trigger (**ST-cell**)

[J. Kulkarni, JSSC'07]



[J. Boley, JLPEA 2012]

- 10T Schmidt trigger (ST) cell: 1.56x increase in read SNM of 6T at 33% area penalty
- 8T asymmetric ST cell: 19% higher RSNM than 10T ST due to single ended reading structure

Assist methods

Write Assist methods

- WL Boost
  - Strengthens passgate device
  - Con: half select instability
- Negative BL
  - Strengthens passgate device
  - Cons: generating negative voltage, increase BL leakage during write

Read assist methods

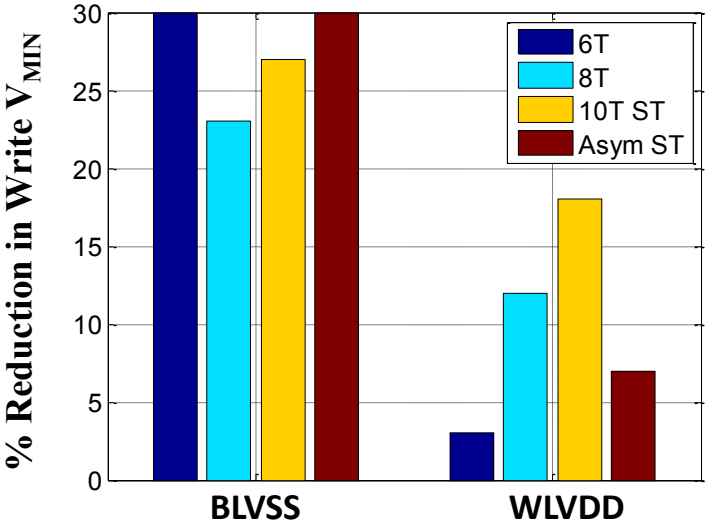
1. WL Boost
   - Increases Ion → faster reads
   - Con: reduces RSNM
2. Negative WL VSS-
   - Decrease Ioff → reduces BL leakage current
3. CVDD boost and CVSS droop
   - Increase the read static noise margin of the cell

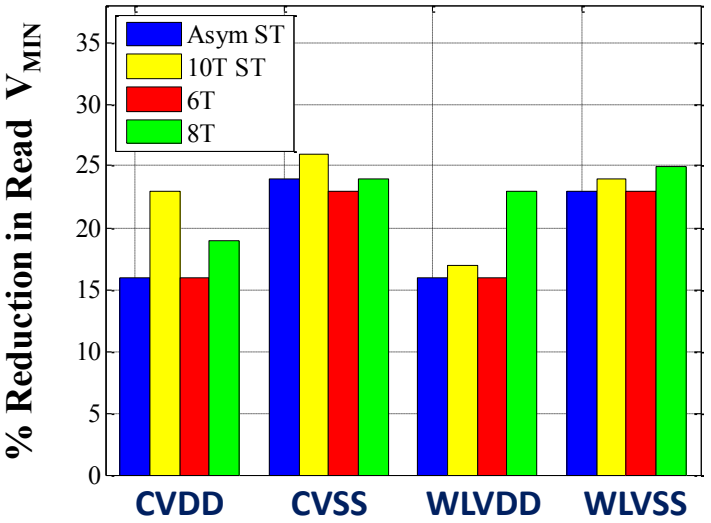
Metrics for Evaluation

1. Reduction in Read and Write V_{MIN} compared to the nominal case (6T bitcell with no assist methods)

Results

- SRAM write limited
  - Write V_{MIN}: 620 mV
  - Read V_{MIN} : 440 mV
- BLVSS reduction: largest reduction of write V_{MIN}
- WL VSS and CVSS: largest improvement on read V_{MIN}
  - Bitline leakage was a major contributor to reduced read margin
- **Impact of assist methods > impact of bitcell topologies on reducing V_{MIN}**





[J. Boley, JLPEA 2012]

Outline

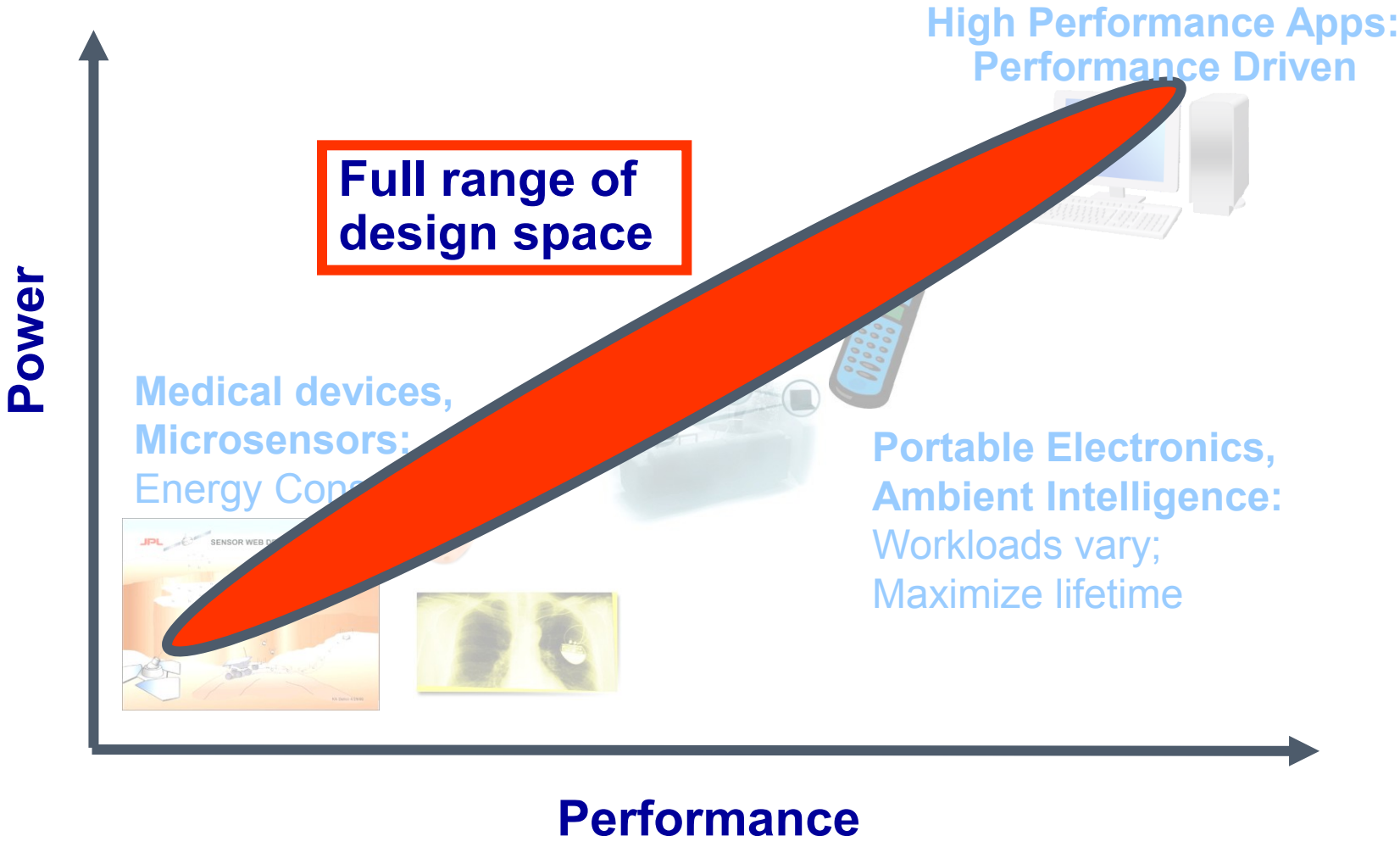
- Motivation
- Sub-threshold SRAM design
- Analysis of sub-threshold bitcell and assist methods
- **A method for evaluating dynamic V_{MIN}**
- Extending the capabilities of ViPro
- A canary based feedback system for PVT tracking
- Schedule

Motivation

- Static stability metrics are optimistic for write
  - Assumes infinite pulse width
  - Doesn't account for transient behavior
  - Upside: shorter simulation times
- Need dynamic metrics for more accurate prediction of V_{MIN}
  - Current metric is T_{CRIT} → the critical or minimum WL pulse width required to write the bitcell
- Scope: Focus on dynamic write-ability

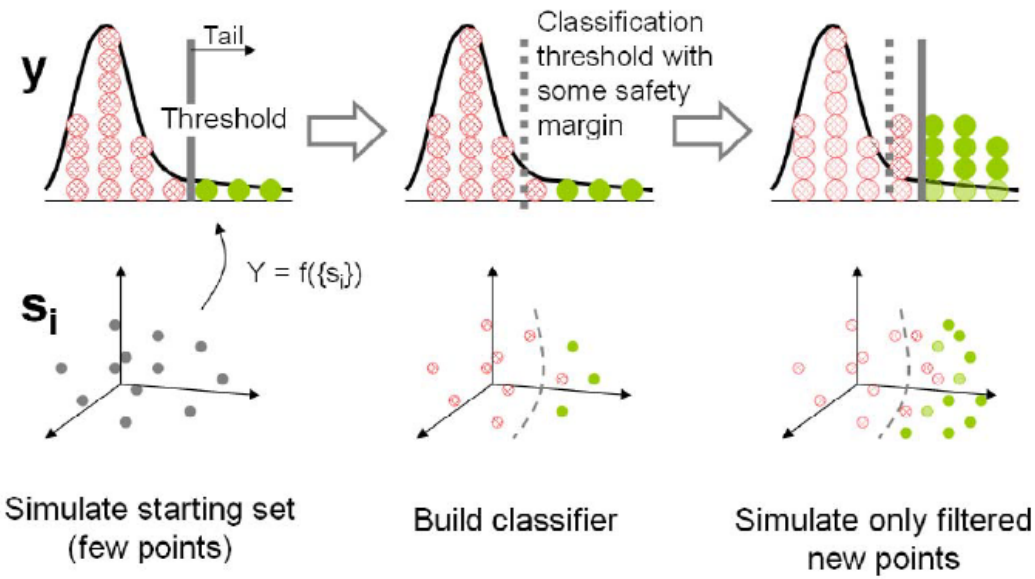
Problem: determining the dynamic write margin of the worst case cell in large (i.e. > 1 Mb) memories requires a prohibitively large number of Monte Carlo (MC) simulations

Scope

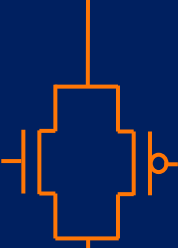

High Performance Apps:
Performance Driven
Full range of
design space
Power
Medical devices,
Microsensors:
Energy Con
Portable Electronics,
Ambient Intelligence:
Workloads vary;
Maximize lifetime
Performance

Prior Art

- Analytical models
  - Less accurate
- Curve fitting
  - Dynamic margin does not closely match any known distribution
- Statistical blockade
  - Run initial simulation
  - Create tail classifier
  - Generate MC points
  - Simulate only samples classified as tail points



[A. Singhee, *DATE, 2007]*



How can we improve statistical blockade?

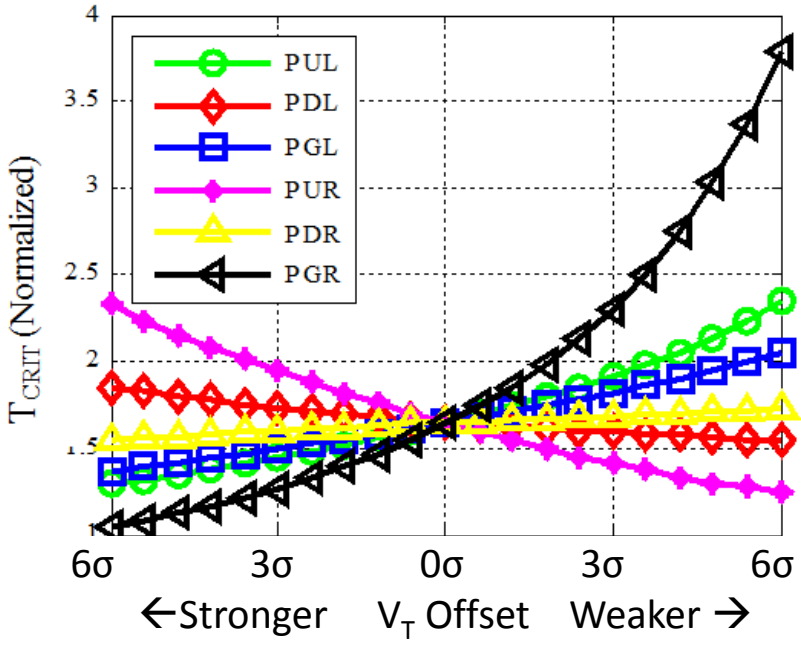
Downsides to statistical blockade:

- Calculating the worst case dynamic margin in a 100 Mb array using recursive statistical blockade requires **over 894,000** total Monte Carlo simulations

Hypothesis: using sensitivity analysis, we can further reduce the time required to calculate dynamic V_{MIN}

Approach: Sensitivity Analysis

Sweep VT of each transistor to calculate $\Delta T_{CRIT} / \Delta V_T$



[J. Boley, DATE 2013]

Bitcell Netlist

↓

Run V_T vs. Tcrit simulation

↓

Polynomial Curve Fitting
$T_{CRIT-OFFSET} = ax^3+bx^2+cx+d,$
$x = V_{T-Shift}$

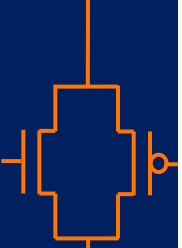
↓

Generate V_T distribution (e.g. in MATLAB)

↓

Compute $T_{CRIT} = T_{CRIT-NOM} + T_{CRIT-PUL-OFFSET} +\ldots+ T_{CRIT-PGR-OFFSET}$

↻ **Repeat for N iterations**



Metrics for Evaluation

1. Speedup gained over statistical blockade (SB)
2. Loss of accuracy compared to SB

Results: Accuracy

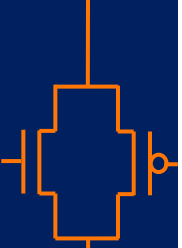
- Comparison between the worst case bitcell as predicted by statistical blockade and sensitivity analysis
- Across VDD, the sensitivity analysis results match closely to the statistical blockade data
- The worst case percent error is 6.83% while the average is ~3%

Modeled data vs. Statistical Blockade (Percentage Error)							
	500 mV	600 mV	700 mV	800 mV	900 mV	1000 mV	Average
100K	6.83	2.96	-0.18	0.83	-4.50	-2.72	3.01
10M	-4.25	-3.69	-2.64	-0.70	0.83	-2.20	2.39
100M	6.51	5.61	4.75	1.21	1.43	-2.27	3.63

Results: Speedup

	Statistical Blockade	Sensitivity Analysis
	Num. simulations	Run Time
Initial Simulation	24,000	18.8 min
100 Kb	107,904	0.72 s
10M	531,096	72 s
100M	231,288	12 min
Total Simulations	894,288	
Total Run Time	60 Hours	32 minutes

Sensitivity analysis provides a 112.5x speed up over recursive statistical blockade with an average percentage error of ~3%



Outline

- Motivation
- Sub-threshold SRAM design
- Analysis of sub-threshold bitcell and assist methods
- A method for evaluating dynamic V_{MIN}
- **Extending the capabilities of ViPro**
- A canary based feedback system for PVT tracking
- Schedule

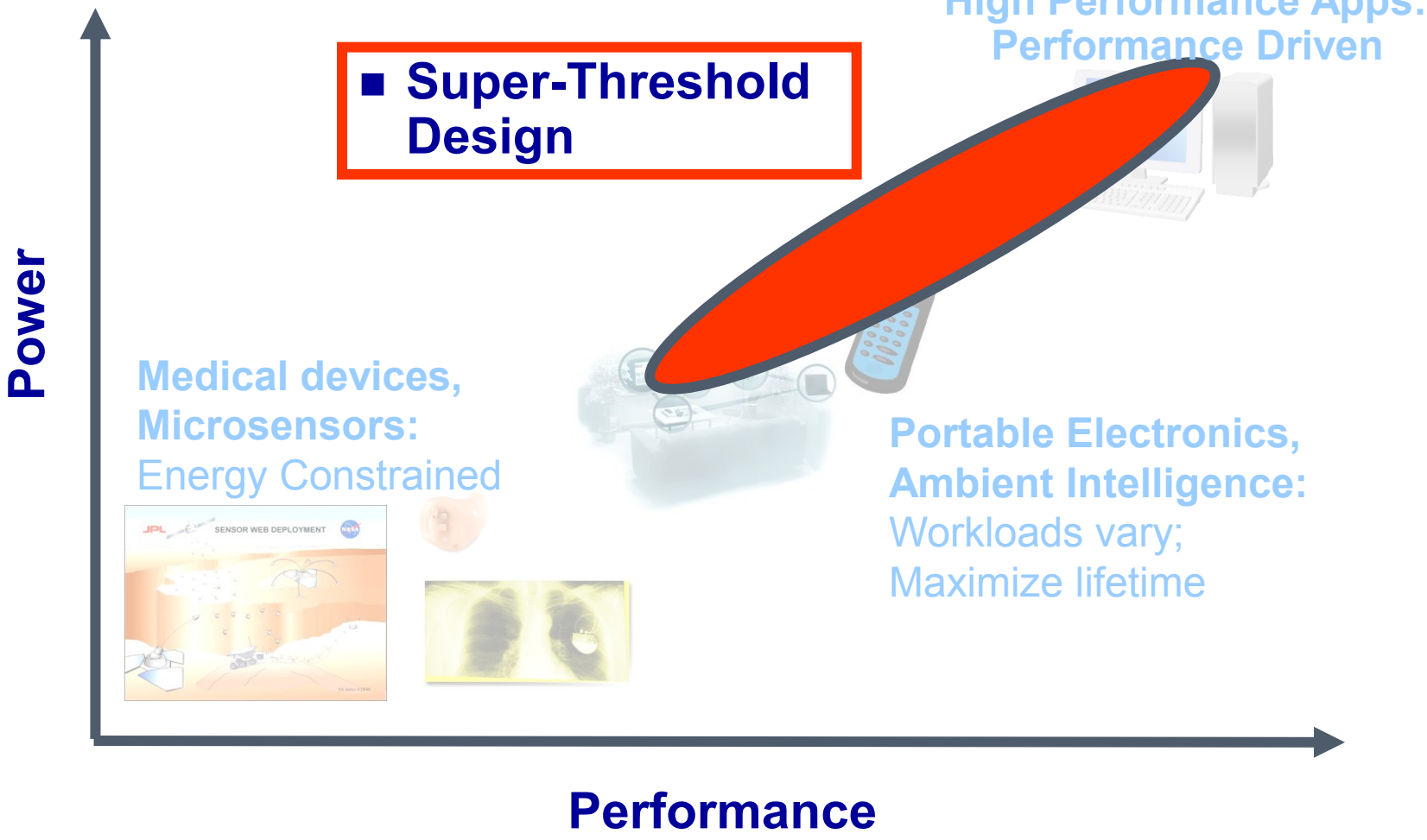
Motivation

- SRAM's face many design challenges in nanoscale technologies:
  - Increased variability
  - Higher leakage
  - Longer interconnect delay
- SRAM limits SoC V_{MIN}
  - 6T bitcell limited by reduced margins
  - Increasing memory capacities
- New knobs have been introduced:
  - Assist methods
  - Alternative bitcells
  - Offset compensated sense amps
- Evaluation of circuit and architectural design decisions in terms of global figures of is difficult without creating a full design

Problem Statement

- Without the proper support structure and tools, it would be nearly impossible to re-design and re-optimize an entire memory by hand every time we try a new circuit, much less explore a technique's impact across different technologies and applications

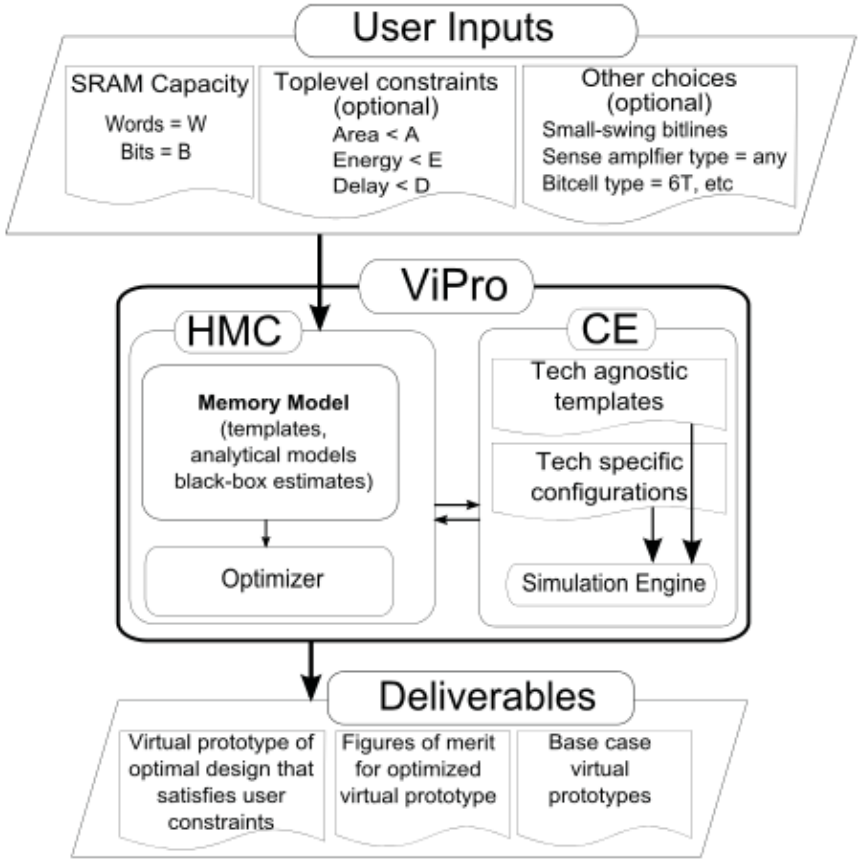
Scope


Power
Performance
Super-Threshold Design
High Performance Apps: Performance Driven
Medical devices, Microsensors: Energy Constrained
Portable Electronics, Ambient Intelligence: Workloads vary; Maximize lifetime
SENSOR WEB DEPLOYMENT

Prior Art

ViPro (developed by Satya Nalam):

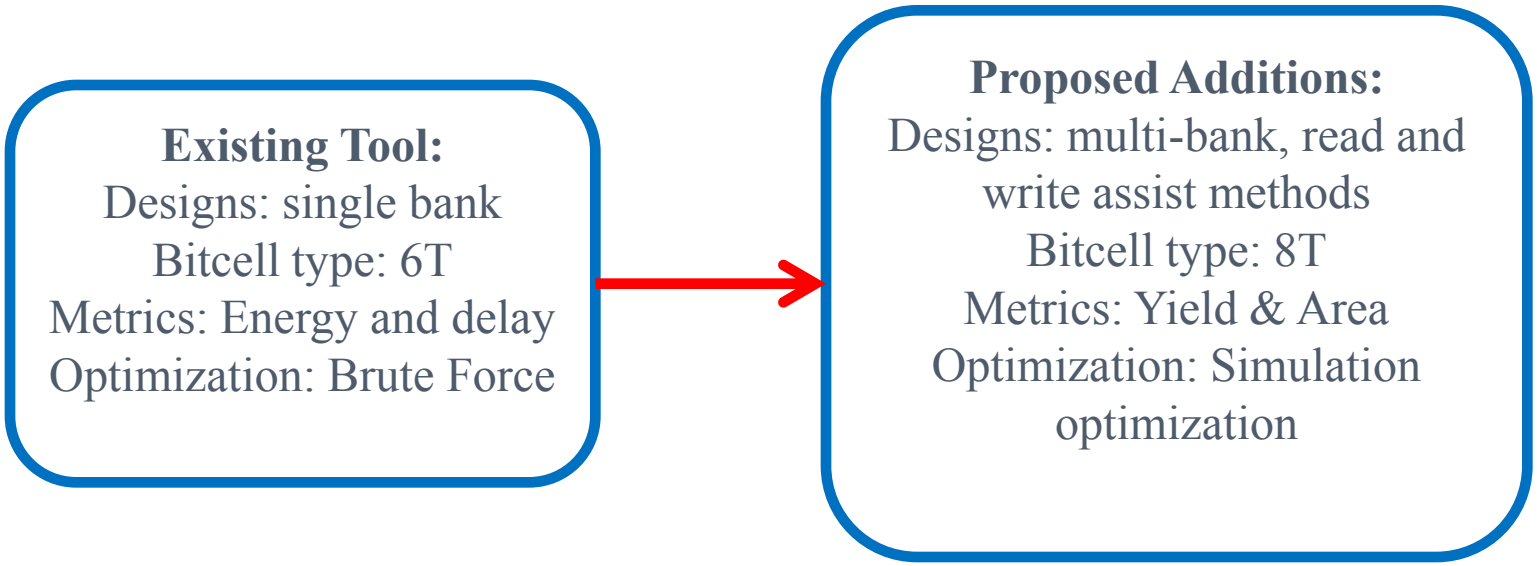
- Tase: technology agnostic simulation environment
  - Uses simulation templates for characterizing SRAMs across technology nodes
- Hierarchical memory model
  - Easily extensible and scalable
  - Faster than running full simulation
- Brute force optimization
  - Supports optimization of a single bank SRAM design
  - Knobs: number of rows and columns
  - Metrics: energy and delay



Hypothesis

- By extending the existing ViPro tool we will be able to explore a much larger design space and run a much larger set of novel experiments which will allow for reductions in SRAM V_{MIN}, increases in yield and a better understanding of design tradeoffs.

Approach Overview



Justifying Our Approach

Multi-bank:

- Single bank designs are limited to <100 Kb
- Current trend →larger capacities

8T:

- Commonly used in L1 cache
- Single ended sensing creates new challenges

Read and write assist methods:

- 6T reaching its fundamental limits due to technology scaling
- Assist methods offer new design tradeoffs
- Choose 2 write assist methods and 2 read assist methods to evaluate

Justifying Our Approach

Yield evaluation:

- Larger designs and higher variability lead to reduced yields
- Clear need to be able to evaluate the effectiveness of new circuit techniques on improving yield
- Use existing techniques for yield evaluation

Area evaluation:

- Key feature of SRAMs is high density

Justifying Our Approach

Circuit and architecture co-optimization:

- Architectural changes affect the optimal circuit level parameters (sizing, number of buffer stages, etc.)
- Brute force only works for small optimization space
- Proposed additions greatly expand the design space
- **Need an optimization engine that can reduce the number of iterations by making smart choices about which knobs to turn**

Evaluation Metrics

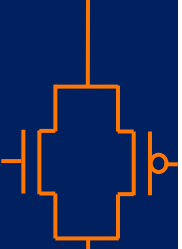
- **Difficult to make direct comparison to existing tools like CACTI**
  - CACTI is an tool for high level design space exploration, however it optimizes at the architecture level only
  - Uses ITRS parameters for evaluating across technologies which may not lead to accurate assessments
- **Proposed metrics for evaluating the tool:**
  - What novel contributions are necessary for creating the tool?
  - What are the novel experiments that the tool enables?
- **Metric for evaluating the optimization engine:**
  - Total speedup compared to brute force optimization

Anticipated Contributions

- Tool for rapid SRAM design space exploration
- Circuit and architecture co-optimization engine

Subset of Anticipated Experiments

- How do assist methods affect read access yield?
- What are the tradeoffs in terms of the global figures of merit between the assist methods?
- What affect does an offset compensated sense amplifiers have on yield and read access time?
- What is the most limiting factor in terms of yield and how does that change with technology scaling?

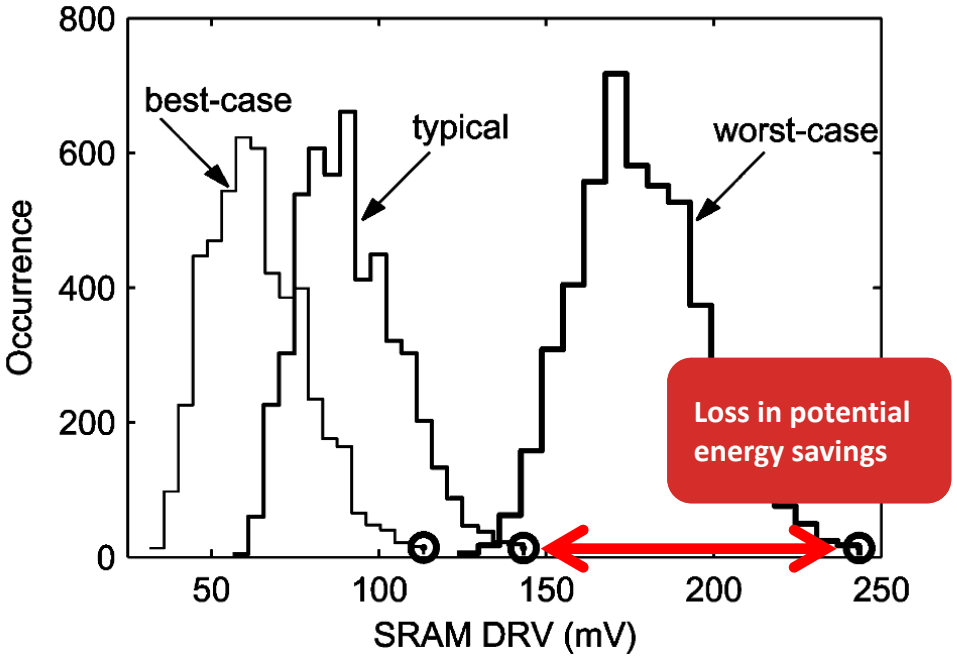


Outline

- Motivation
- Sub-threshold SRAM design
- Analysis of sub-threshold bitcell and assist methods
- A method for evaluating dynamic V_{MIN}
- Extending the capabilities of ViPro
- **A canary based feedback system for PVT tracking**
- Schedule

Motivation

- Typical commercial designs must guard band for a range of process, temperature, voltage (PVT) corners
- Guard banding must be done pre-fabrication to ensure reliability across all corners

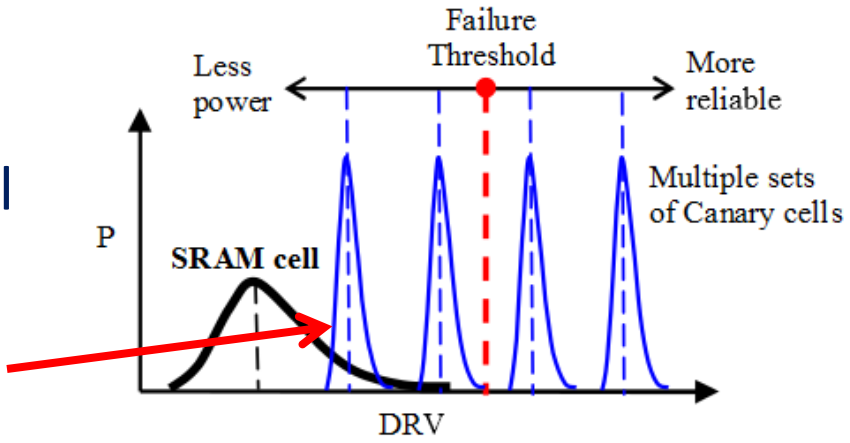


[Wang, JSSC 2008]

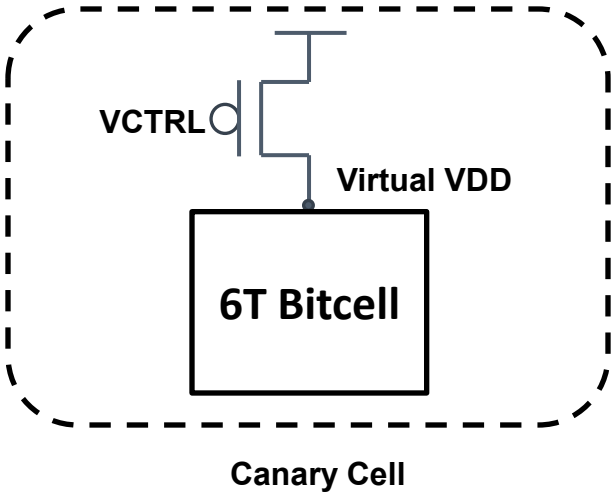
Problem: Conservative guard banding eliminates potential post-fabrication energy savings.

Prior Art- Canary Feedback System

- Previously applied to data retention voltage (DRV)
- Canary cells designed to fail at regular intervals
- **Canary cells use upsized transistors to reduce local mismatch**
- Closed loop controller detects canary failures
- Direct tradeoff made between reliability and power



[Wang, JSSC 2008]



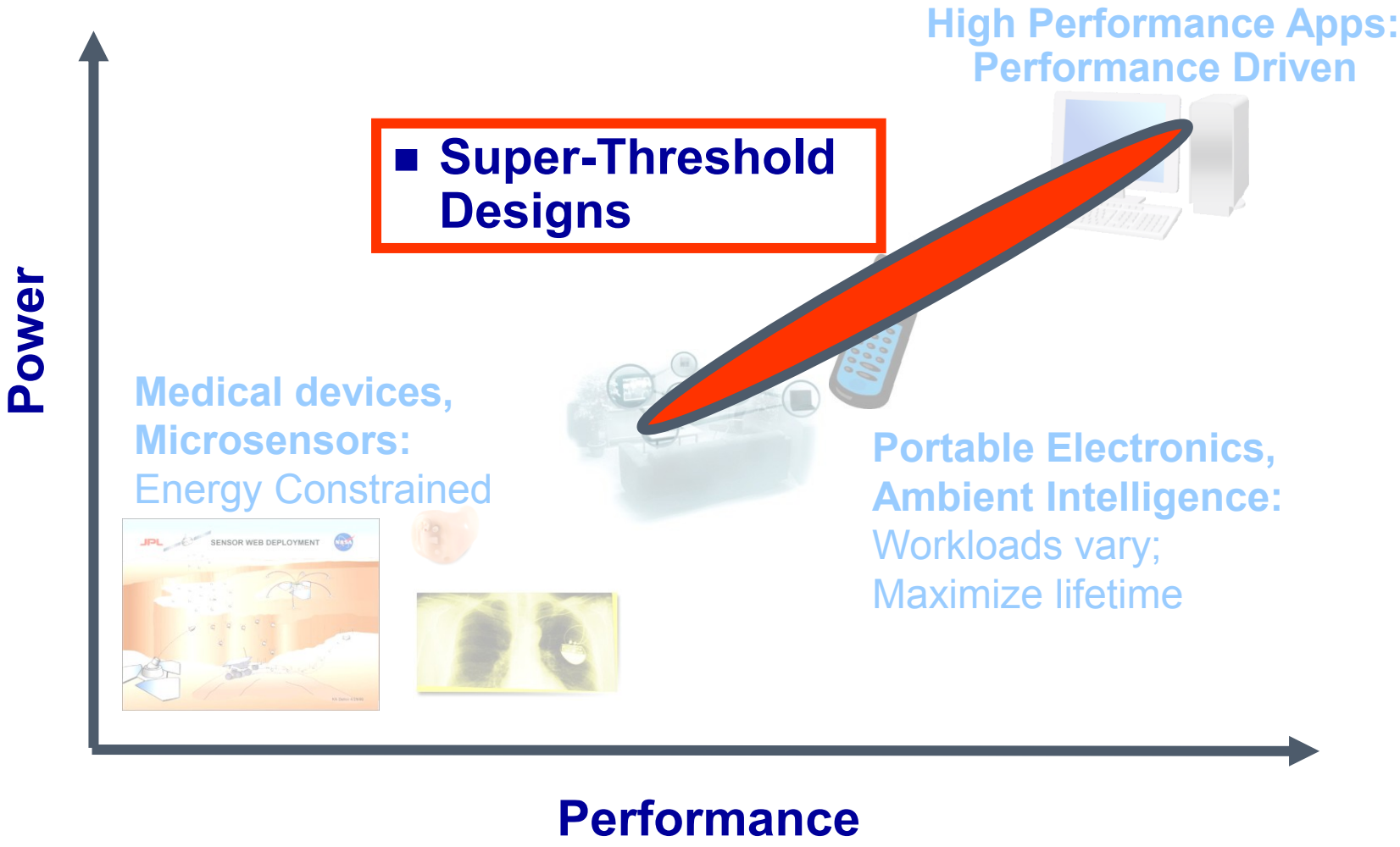
Previously reported: 30x power savings over conservation guard banding [Wang, JSSC 2008]

Hypothesis

- We can achieve significant power savings over conservative guard banding by using a canary based feedback system to reduce write V_{MIN}

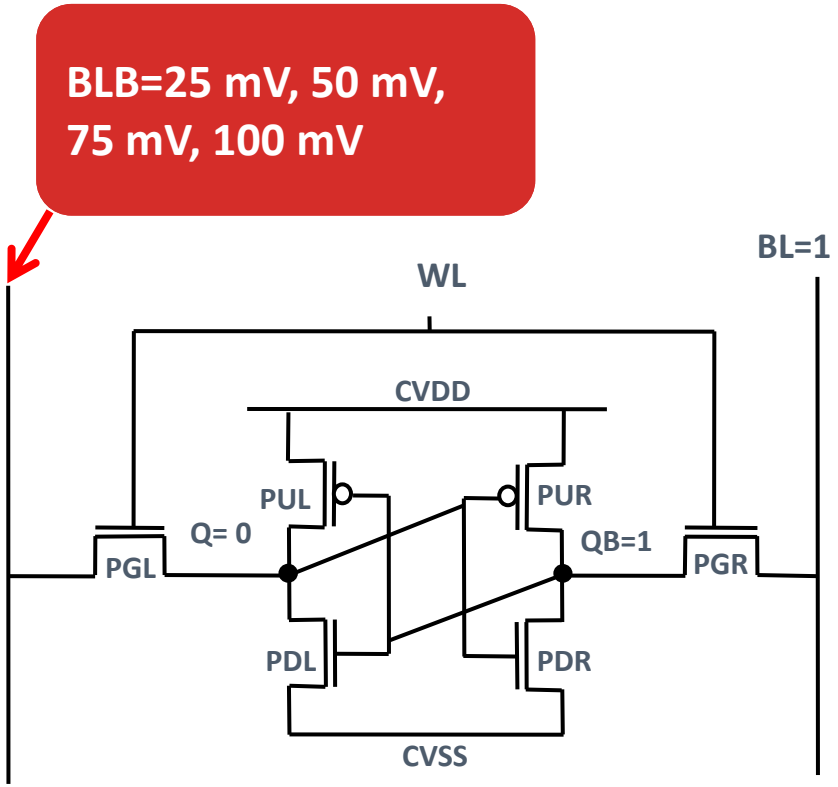
NOTE: While a full canary system would need to also consider the read path, we have chosen to limit our scope to only the write operation in this work.

Scope

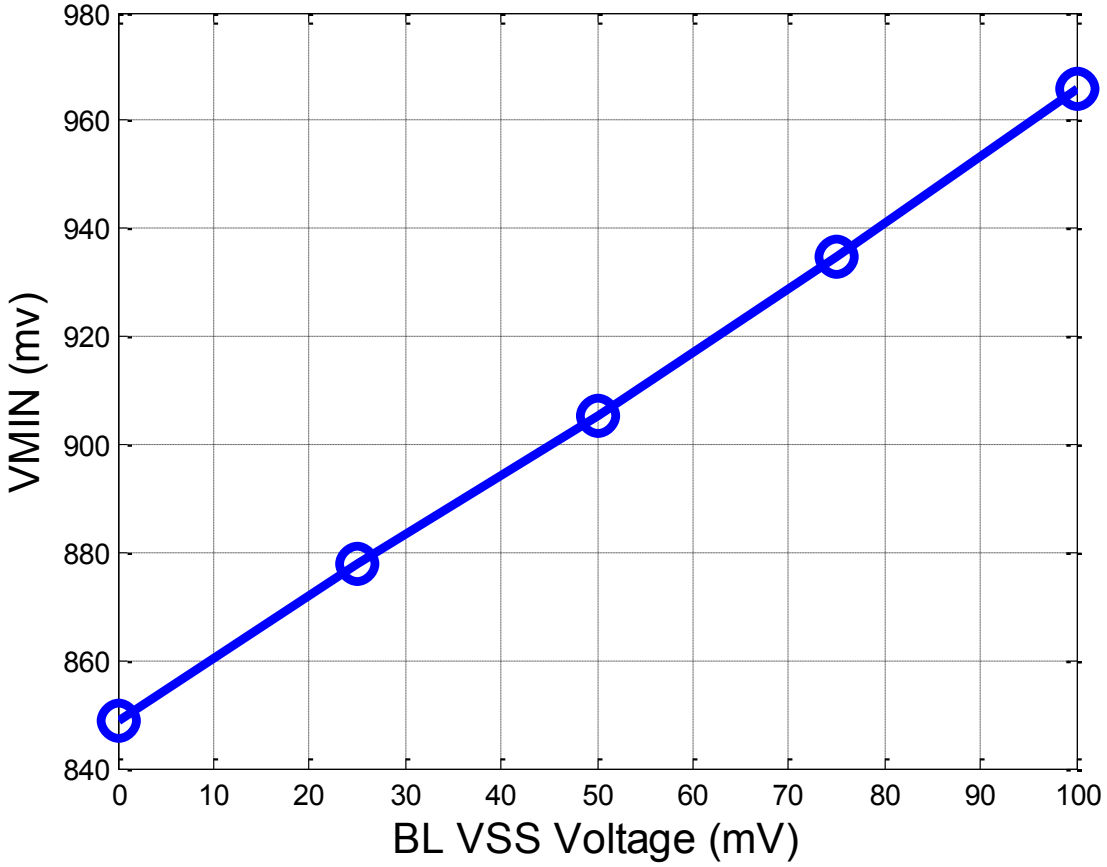

High Performance Apps:
Performance Driven
Super-Threshold Designs
Power
Medical devices,
Microsensors:
Energy Constrained
Portable Electronics,
Ambient Intelligence:
Workloads vary;
Maximize lifetime
SENSOR WEB DEPLOYMENT
Performance

Approach: Canary Design

- Boost BL V_{SS} above ground
- Keep WL pulse width equal to core array
- Lower voltage until failures are detected
- **Challenge: setting the BLB voltage**



Preliminary Testing



Approach: Other Canary Designs to Consider

Reverse Assist Methods:

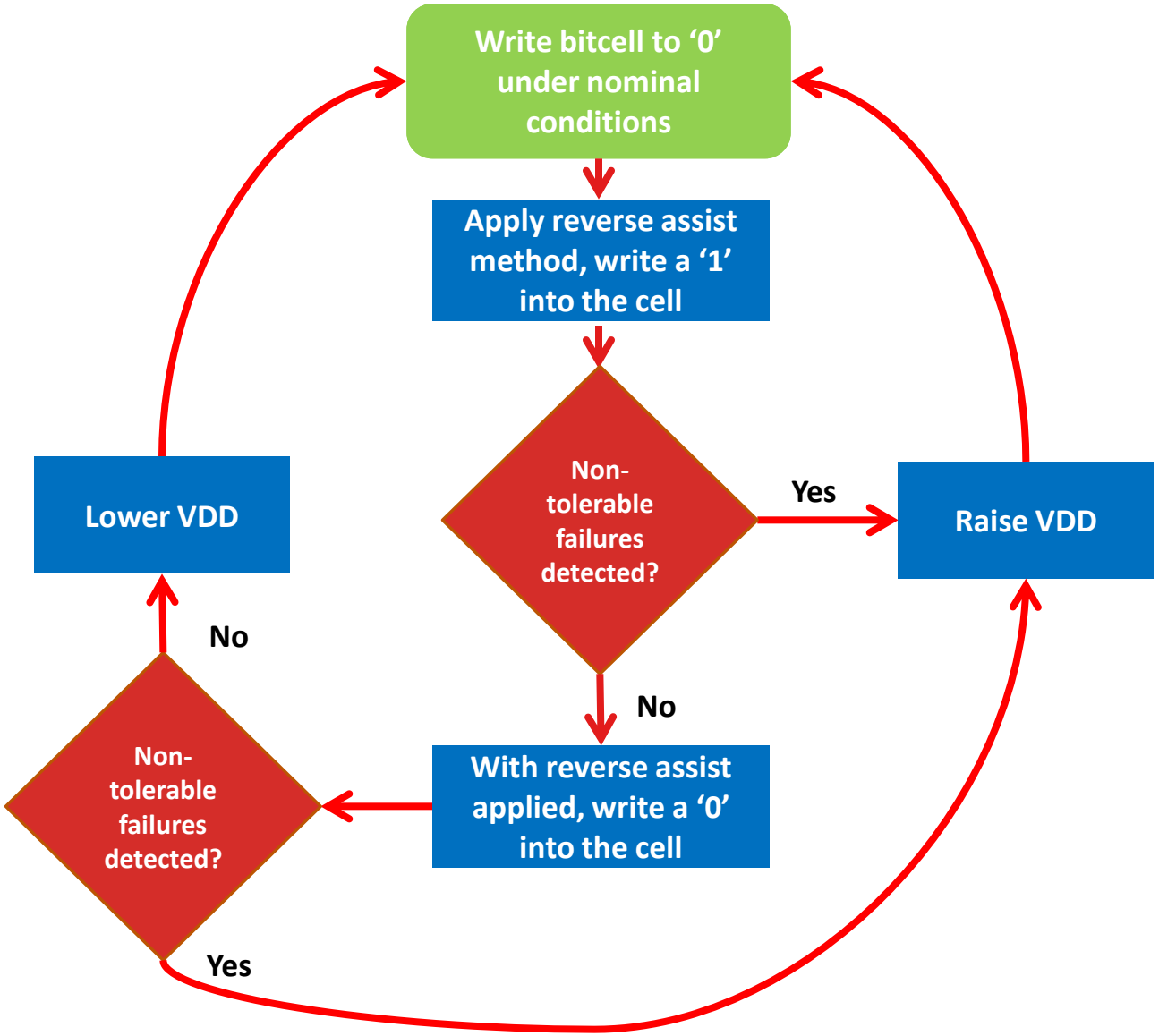
- **Droop WL VDD:**
  - Reduce the strength of the passgate transistors
- **Negative CVSS:**
  - Increase the strength of the cross coupled inverters
- **Boosted CVDD:**
  - Increase the strength of the cross coupled inverters

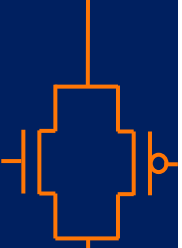
Shorten write wordline (WL) pulse width:

- Shorter pulse width → higher failure voltage

Scope: choose one or two of these methods for implementation on a test chip

Approach: Control Loop





Anticipated Contributions

- Canary design that tracks PVT variation and fails at specified intervals
- Control system that can regain energy lost due to conservative guard banding by further lowering write V_{MIN}

Schedule

Project	Milestone	Status/Target
BSN Sub-VT SRAM	Rev. 1- Design/layout/testing	*Completed*
	Rev. 2- Design/layout	*Completed*
	Rev 2: Chip Testing	August-2013
Dynamic Write V_{MIN} Estimation	Create/verify model	*Completed*
Sub-threshold bitcell analysis	Test Chip	*Completed*
Virtual Prototyping Tool	Support multi-bank design	*Completed*
	Support for 8T bitcell	*Completed*
	Verify Model Accuracy	July-2013
	Integrate two read and two write assist features	September-2013
	Integrate area estimation	December 2013
	Integrate yield estimation	April 2014
	Optimize using simulation optimization algorithm	August-2014
Canary Feedback System	Evaluation of canary design	December 2013
	Design of voltage control loop	February 2014
	Simulation/Verification/Layout	April 2014
	Chip Testing	December 2014
Write up	Thesis Writing	February-2015

Publications

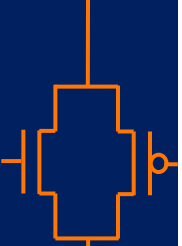
1. F. Zhang, Y. Zhang, J. Silver, Y. Shakhsheer, M. Nagaraju, A. Klinefelter, J. Pandey**, J. Boley**, E. Carlson, A. Shrivastava, B. Otis, and B. H. Calhoun, "A Battery-less 19µW MICS/ISM-Band Energy Harvesting Body Area Sensor Node SoC," *ISSCC*, February 2012.
2. **J. Boley**, J. Wang, and B. H. Calhoun, "Analyzing Sub-Threshold Bitcell Topologies and the Effects of Assist Methods on SRAM V_{MIN}", *JLPEA*, April 2012.
3. Y. Zhang, F. Zhang, Y. Shakhsheer, J. Silver, A. Klinefelter, M. Nagaraju, **J. Boley**, J. N. Pandey, A. Shrivastava, E. J. Carlson, A. Wood, B. H. Calhoun, and B. Otis, "A Batteryless 19 µW MICS/ISM-Band Energy Harvesting Body Sensor Node SoC for ExG Applications," *JSSC,* 2013.
4. **J. Boley**, V. Chandra, R. Aitken, and B. Calhoun, "Leveraging Sensitivity Analysis for Fast, Accurate Estimation of SRAM Dynamic Write V_{MIN}," *DATE*, 2013.
5. **J. Boley**, P. Beshay, B. Calhoun, "Virtual Prototyping (ViPro) Tool for Memory Subsystem Design Exploration and Optimization," TECHCON, 2013.

Anticipated publications:

1. SRAM Circuit and Architecture Co-Optimization
2. Subthreshold SRAM Design Featuring Low Energy Read operation
3. Evaluation of SRAM Assist Methods on Top Level Design Metrics
4. Optimization of SRAMs for Improved Yield
5. Using simulation optimization for SRAM design space exploration
6. Canary based closed-loop control system for optimizing write V_{DD}

Acknowledgemetnts

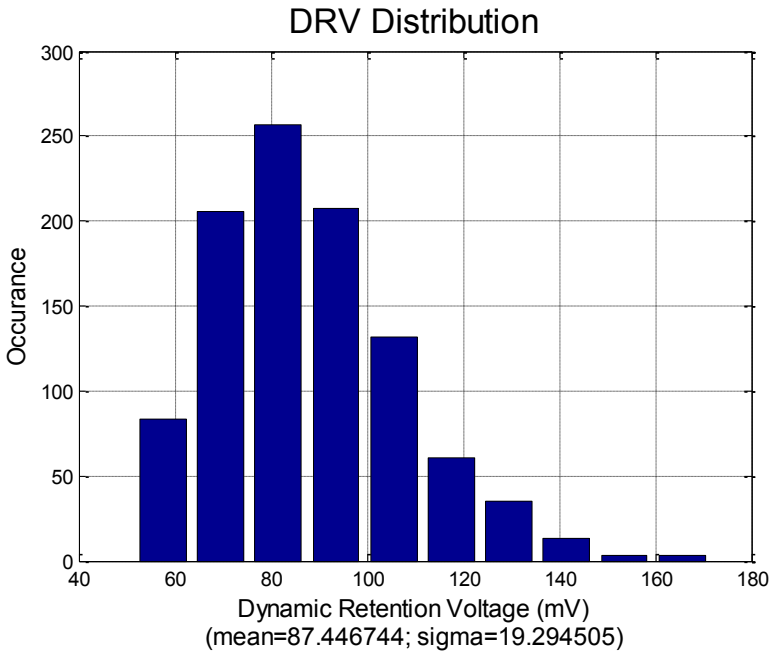
- Committee Members
  - Ben Calhoun, Adviser
  - Joanne Dugan, Chair
  - Don Brown, Member
  - John Lach, Member
  - Ronald Williams, Member
- Fellow Students
  - Randy, Satya, Jiajing, Sudhanshu, Joe, Yousef, Kyle, Yanqing, Aatmesh, Alicia, Seyi, Peter, Arijit, Chu, Patricia, Divya, Yu, He
- ARM Research
  - Vikas Chandra and Rob Aitken



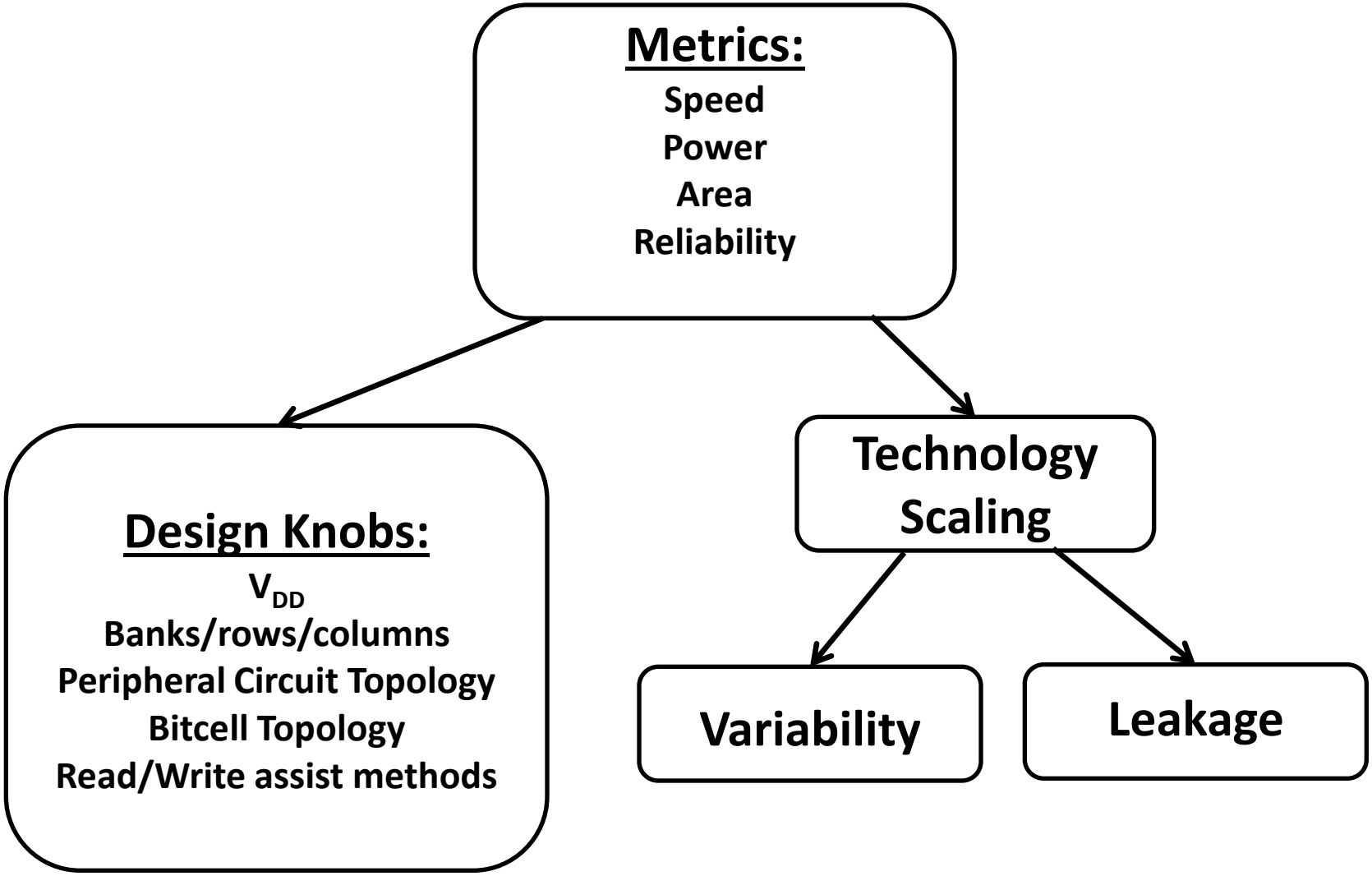
Backup slides

Key Challenges: Estimating Yield

- Monte Carlo (MC) = Gold Standard
- >1 MB arrays make MC sims infeasible
- Distributions not always normal

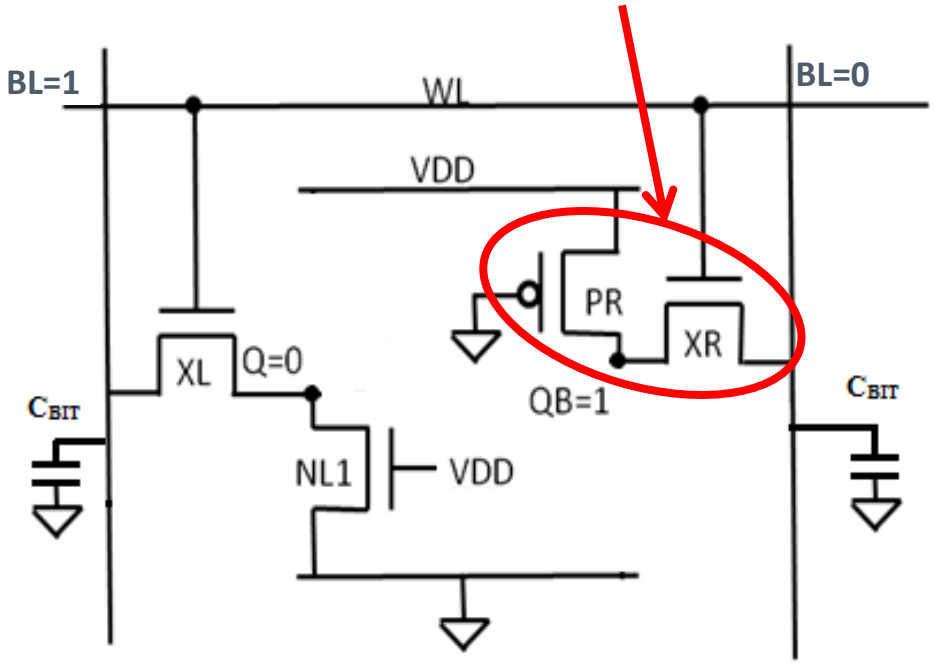


Key Challenges: Evaluating Design Decisions



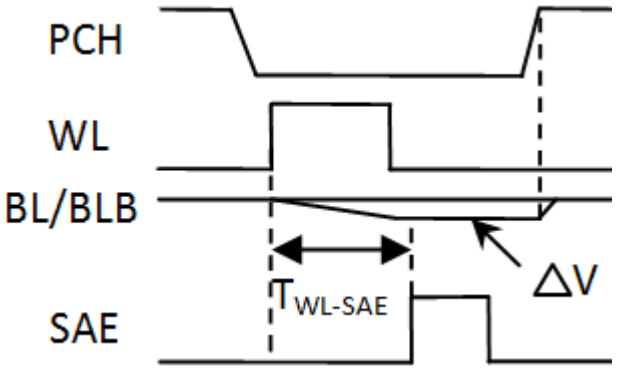
Key Challenges: Reduced Write-Ability

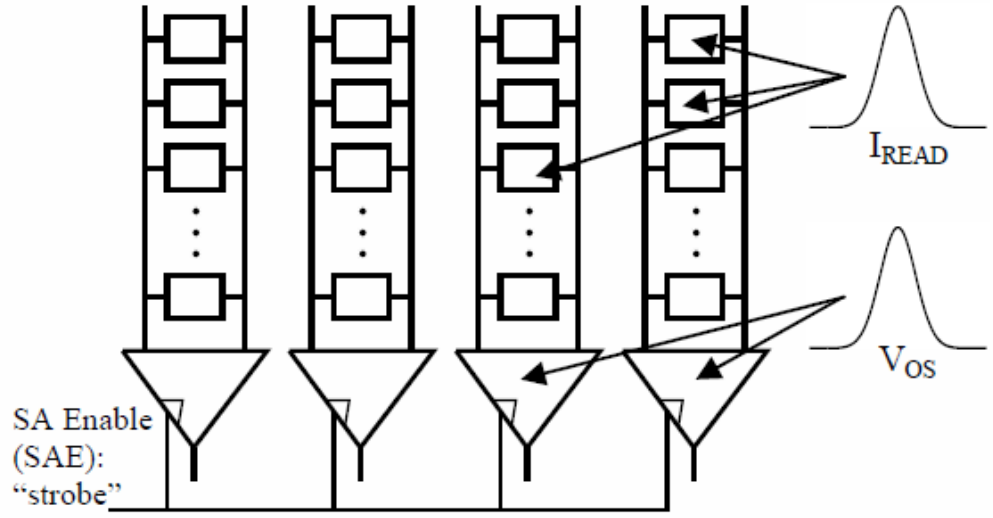
- Ratioed design
- PR minimum sized



Ratioed circuit design → higher sensitivity to V_T variation

Key Challenges: Read Access Stability





[J. Ryan, ISLPED 2011]

- Goal: minimize T_{WL-SAE}
- I_{READ} determines ΔV
- V_{OS} → input referred offset of the sense amp (SA)
- Increased leakage also leads to reduced ΔV
- **Failure point: ΔV < V_{OS}**

Motivation of Sub-threshold (V_{DD}<V_T)

- Sub-threshold benefits: V_{DD} from [1.8,1.0]V to [0.4,0.2]V

Leakage Power Decreases: Power = V_{DD} I_{off}

V_{DD} goes down: 2.5X to 9X

DIBL reduces $I_{sub\text{-}threshold}$: 2X to 10X

I_{gate} and I_{GIDL} become negligible

Pleak: 5X to 90X

Energy Consumption Decreases

$E_{active} = CV_{DD}^2$

E_{total}/operation minimized in sub-V_T

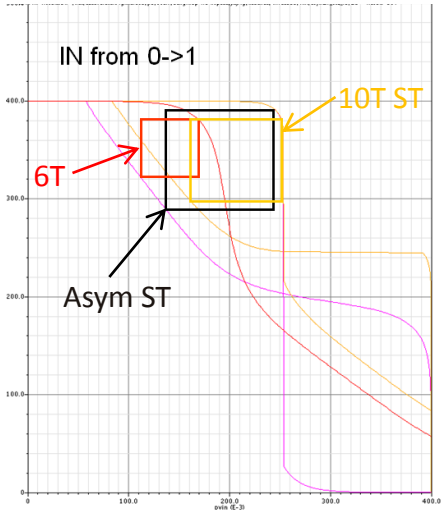
Aging Effects Improve

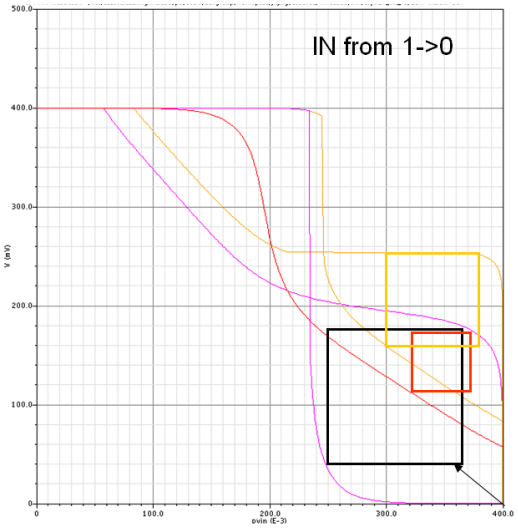
NBTI, EM, TDDB

Main Limitations: Variation, Slow Speed

Approach: 8T Asymmetric Schmitt Trigger Bitcell

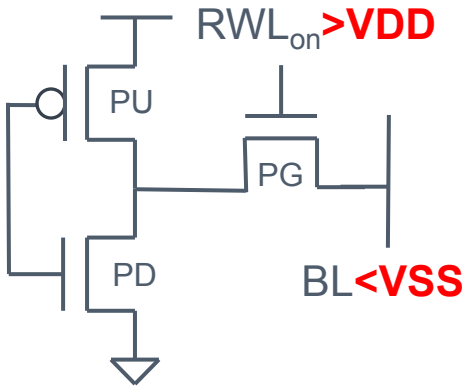
- Uses single-ended reading and assymetric inverters similar to the 5T cell described in [Nalam, CICC'09] to increase read margin
- Write operation similar to 6T write
- Asymmetric ST cell achieves 86% higher static read noise margin (RSNM) than the 6T cell, and 19% higher RSNM than the 10T ST cell





Approach: Write Assist Methods

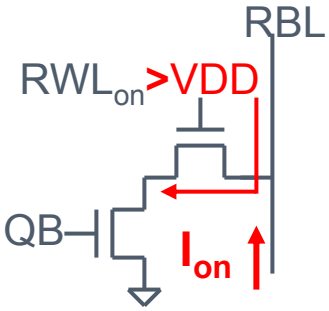
- Goal
  - Weaken pull-up FET
  - Strengthen pass-gate FET
- Knobs
  - Size pass-gate to pull-up ratio (not efficient)
  - Collapse V_{DD} to weaken PFET
  - Boost WL V_{DD}
    - Cons: half selected cell stability
  - Reduce BLVSS
    - Cons: increased BL leakage

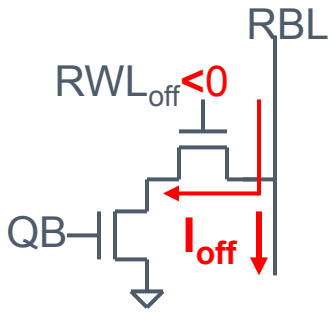


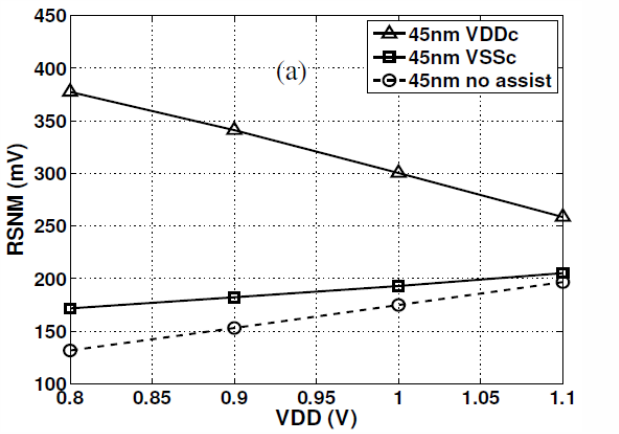
VGS_{PG}>VDD

Approach: Read Assist methods

- Keys
  - Increase I_{on}
  - Reduce I_{off} (BL leakage current in unaccessed cells)
- Knobs







[R. Mann, ISQED'10]

A: boosted on-WL

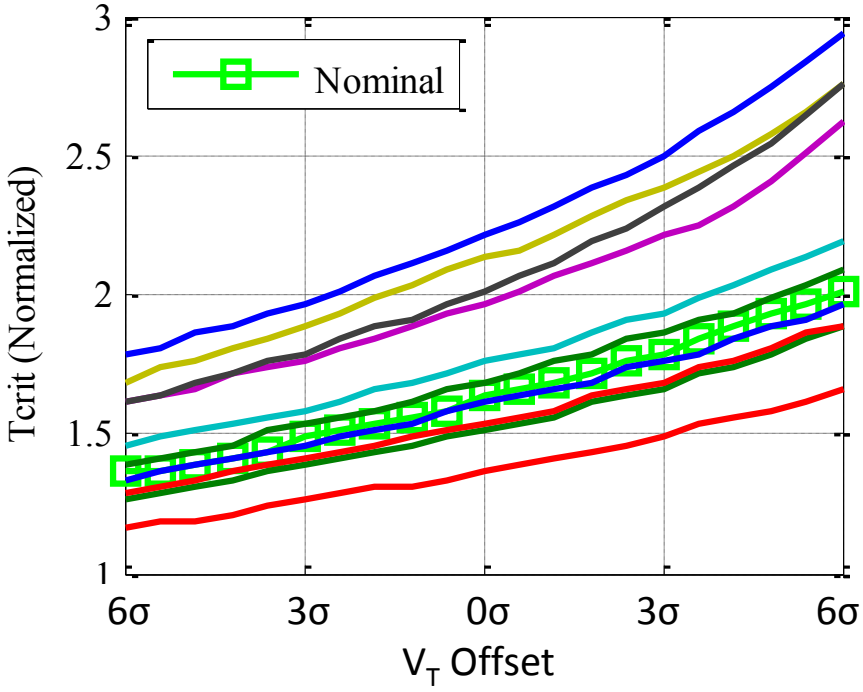
B: negative off-WL

C. boosted bitcell voltage

D. negative bitcell VSS

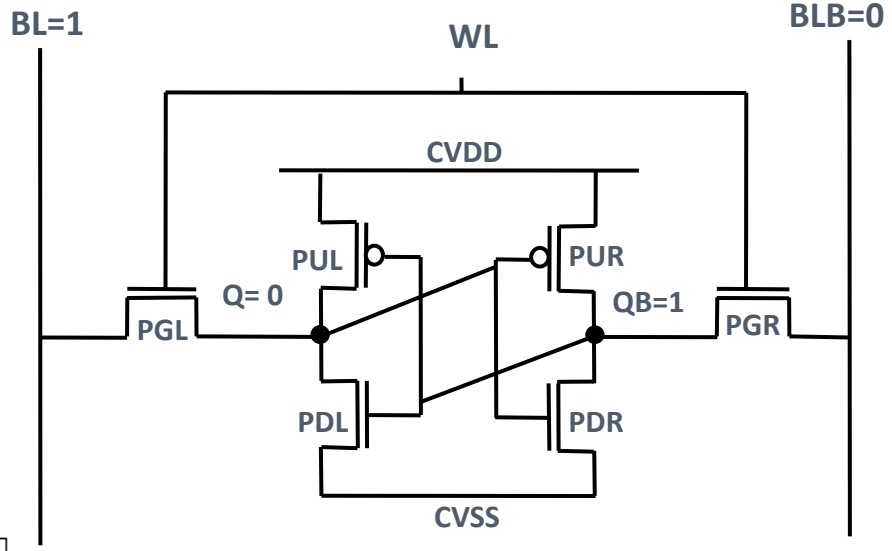
Approach: Verifying Assumptions

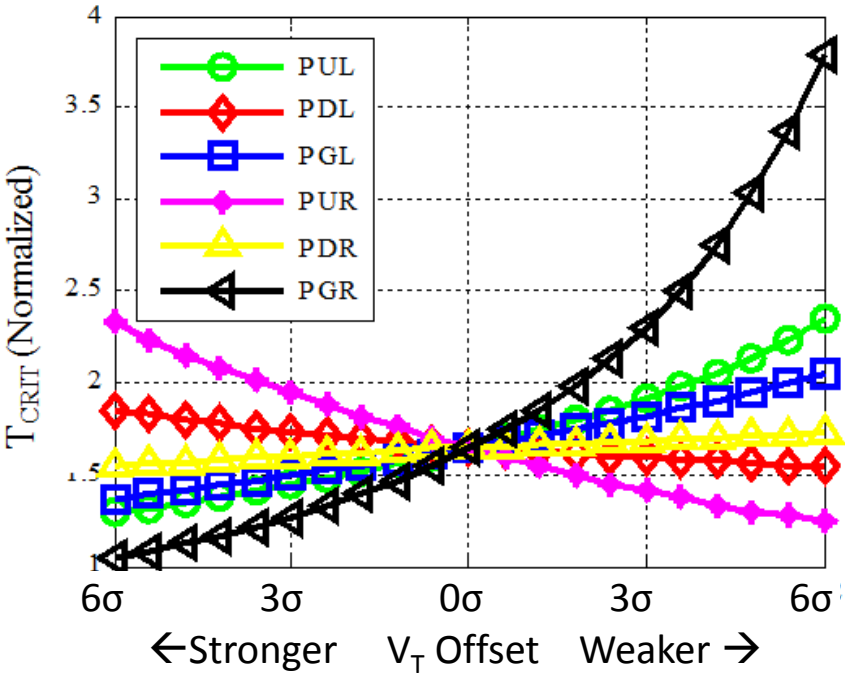
- Experimental setup:
  - Add variation to other five transistors, sweep V_T of single transistor
  - Repeat for each transistor
- Expected output:
  - Shape of the sensitivity curve unchanged
  - Nominal value (0σ) shifted higher or lower



Approach: Sensitivity Analysis

- Sweep VT of each transistor to determine sensitivity
- Generating the V_T curves requires only 1080 simulations
- Once the V_T curves have been generated, the Monte Carlo data can be run through the model, and the worst case bitcell can be quickly found

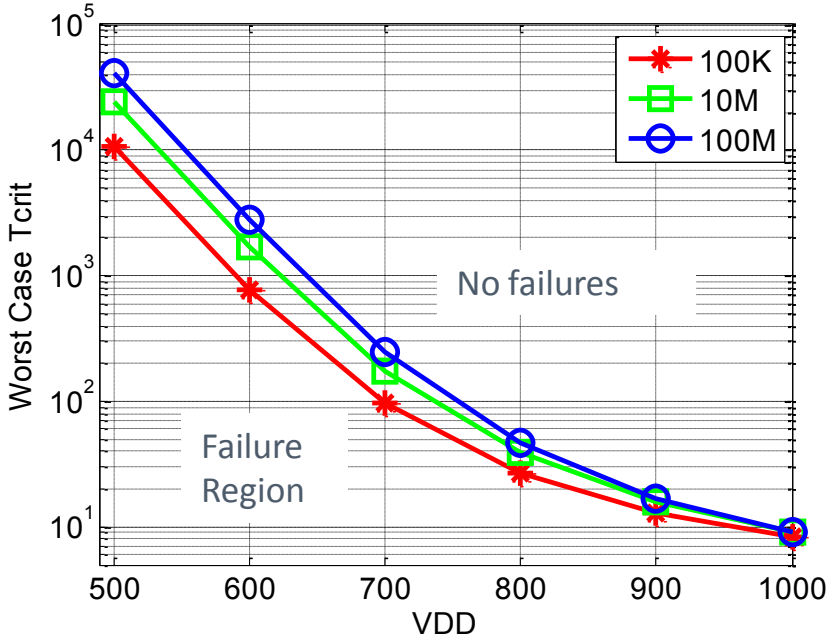




Use sensitivity analysis to calculate $\Delta T_{CRIT} / \Delta V_T$ for each transistor

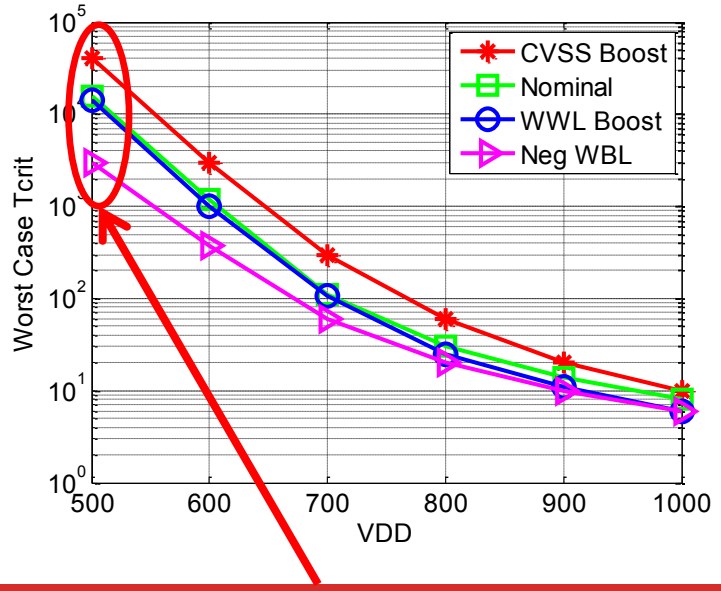
Applying Results to Dynamic V_{MIN}

- In order to calculate V_{MIN} we can repeat the procedure for varying VDD
- We chose 6 points from 0.5-1V
- The plot shows the worst case bitcell for a given VDD, varying the array size
- The curve represents the point of single bit failure, below the curve represents multiple failures, above the curve represents no failures



Applying Results to Assist Methods

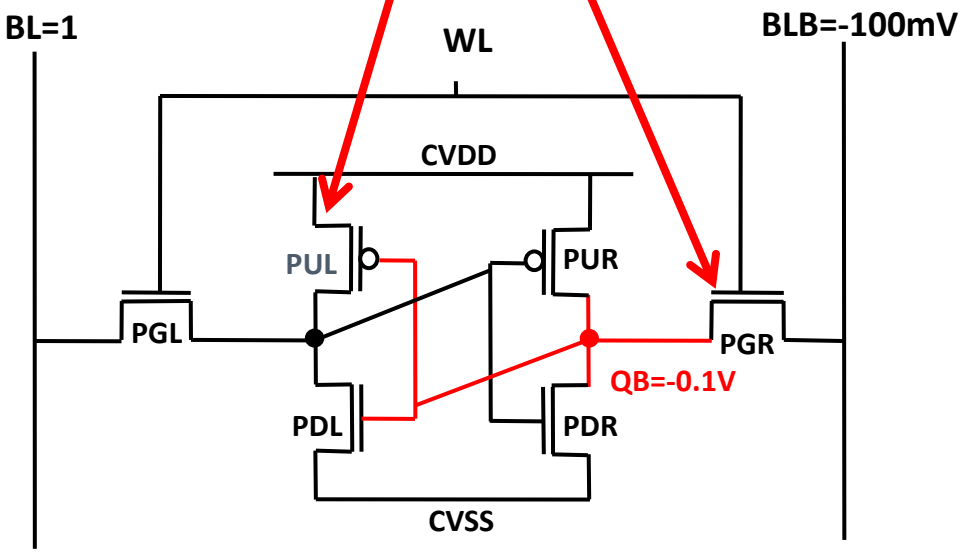
- Memory size is 1 Mb
- Lower VDD shows negative WBL has clear advantage
- ΔV = 100 mV for each assist method



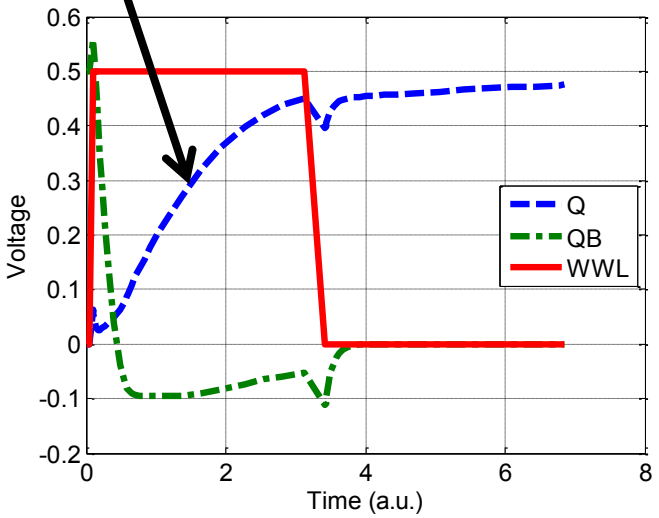
Negative BL reduction reduces the worst case T_{CRIT} close to an order of magnitude more than WWL boosting at 500 mV

Advantage of Negative BL Reduction

Negative BL strengthens pass-gate and pull-up devices



At 500 mV, the majority of write time is spent pulling Q high



Negative BL passes a stronger '0' into the cell, effectively strengthening the PUL transistor and decreasing the write time. This gives negative BL an advantage over WWL boosting at lower VDD.

Project Summary